Op Amp Design in Nanoscale Processes Using Fixed-Length Devices

by

Daniel Saari

A thesis
presented to the University of Waterloo
in fulfillment of the
thesis requirement for the degree of
Master of Applied Science
in
Electrical and Computer Engineering

Waterloo, Ontario, Canada, 2014

Author's Declaration

I hereby declare that I am the sole author of this thesis. This is a true copy of the thesis, including any required final revisions, as accepted by my examiners.

I understand that my thesis may be made electronically available to the public.

Abstract

Analog integrated circuit design has become increasingly difficult in modern fabrication processes. The motivation for digital speed has posed problems for mixed-signal projects that wish to implement digital and analog blocks on the same chip. With the introduction of multigate transistors (also known as FinFETs), the challenges for analog design increase. This is due to the fact that FinFET devices will no longer have a continuum of width and lengths sizes (as previous technologies have exhibited), but instead, these parameters are now quantized. This work proposes a potential solution to the fixed-length problem, in a topology termed the "series-stack".

Foundries plan to launch the FinFET technology with a number of fixed-sized transistors (typically with minimum length). To the digital designer, this poses little problem, but for analog circuits, not being able to control device length compromises the ability to meet gain specifications. This work explores a simple method for implementing longer devices: connecting transistors in series, herein called series-stack. To test the feasibility of this architecture, a two-stage CMOS operational amplifier is designed. In lieu of application-specific design constraints, a structure strategy is presented. A key motivation for the series-stack as well as the design strategy is to bring the analog design process up a level of abstraction. The amplifier was planned to be put through the entire design cycle, from conception to lab testing, giving insight into the accuracy of simulation models.

Schematic and post-layout results were collected from the TSMC 65nm kit. Analysis of the results yield obvious simulation discrepancies. Namely, the schematic simulation vastly overestimates the parasitic resistances and capacitances when using finger-gate techniques. This is an important problem for which possible solutions are discussed. Additionally, the results show significant differences between conventional bulk length and series-stack, with a relative error spanning from 2% to 20% depending on the performance metric. Yet, most discrepancies are expected, and the two implementations follow similar trends with respect to current density and length.

A final verdict cannot be delivered until physical chip testing is conducted, which is left to future work (complications in timeline did not allow for the lab test results to be included). Although chip testing was not completed, a thorough testing plan is formulated. Despite physical testing, the series-stack is deemed a suitable alternative to long transistor designs, especially when considering the organizational advantages at the layout level.

Acknowledgements

Too many individuals have aided me through my academic journey; it would simply take too long to properly thank all those involved. But, there are some who must be explicitly named to show my sincere gratitude.

I would like to thank Professor Nairn for being a stellar supervisor. There is not a day that goes by where I do not consider myself lucky for getting an opportunity to work with someone as knowledgeable and encouraging. He has shaped my understanding of circuits and engineering, while always providing invaluable insight into the big picture. I thank you for taking a chance with me, and displaying patience throughout this process. I am also greatly honoured by my two thesis readers, Professor Adel Sedra and Professor Vincent Gaudet, who have had an enormous impact on my understanding of electronics. I thank them for their comments and insight.

My fellow colleagues have been instrumental in the completion of this work. Without Adam Neale, this work would have never been fabricated; his time and patience are sincerely appreciated. In addition, I thank Pierce Chuang who helped with understanding the TSMC 65nm kit, as well as Phil Regier. Sriram Moorthy put up with many of my ramblings concerning circuits, I thank him. I also thank Adam Bray, who, in a short amount of time, provided me with essential information regarding analog design flow.

I am fortunate to have a great emotional support staff. Mom, dad, Korey, Michelle, Laura, Jay, Mark and May, thank you for being ever supportive of my ambitions, and always believing in me. Mom, thank you for the endless discussions regarding my career interests (I have finally found my passion), and thank you for showing me the value of hard work. To my entire family, words cannot express my gratitude. I would also like thank God for His many blessings.

Lastly, I would like to thank Laura, who has experienced the stress of chip tape-out, without ever being involved in circuits. Thank you for being the calming voice, the encouraging coach, and the uplifting teammate. I appreciate your curiosity regarding all my odd life wisdom; anyone else would think it is nonsense.

Dedication

This thesis is dedicated to my grandmother, Rita.

Table of Contents

List of Tables

List of Figures

Chapter 1

Introduction

1.1 Motivation

Due to increases in abstraction and the discretization of signals, digital circuits continue to follow Moore's Law of scaling.[1] Analog circuits have a lack of impetus to follow this trend, since their performance specifications typically require large devices. Hence, moving to higher resolution processes is not efficient from a cost perspective. In addition, design becomes unpredictable due to break down in models used to characterize the device behaviour. Yet, high performance systems require the digital and analog domains to be in close proximity. This is not a significant concern, since it is always possible to implement larger devices. However, going below a resolution of 20 nm sets a constraint on the device size, so analog circuits will have a problem building large devices. This work aims to find a solution to this problem.

To fully understand the principles discussed, some background in device physics and electronics is required. A brief review of relevant topics is discussed in Chapter 1, while the problem is formally outlined in Chapter 2. This work's solution is presented in Chapter 3, with the following chapters aiming to evaluate the feasibility.

[1] Moore's Law outlines the trend of device density, where the number of devices per area doubles approximately every two years.

1.2 Analog Signals

Over the past century, human beings have begun to connect on a global scale. Starting with the invention of aviation (and before that naval technology), the world has seemingly shrunk. Mobile phones only accelerated the fire of mankind to always be connected and have communication with others where distance does not seem to be a significant factor. Telecommunications, the internet and, if one is so bold, technology as a whole has been founded and fuelled by the ability to control electricity; without this, there would be no mobile phones, Internet or computer systems (at least not to today's scale, one could build it with water - but that would make an awful mess).

In order to use electricity, it must be possible to transfer information from mother nature's chosen medium to an electrical signal. To do this, transducers are used to take the information from its natural state (sound-waves, electromagnetic waves, physical vibrations, temperature, etc.) and convert it to the electrical domain. Due to the design of transducers, these electrical signals are now analogous to the natural phenomena. For example, a microphone reacts to the sound wave's tones and loudness, converting them to an electrical signal that has an analogous frequency (tone) and amplitude (loudness). Therefore, these electrical representations of the physical world are called analog signals.

As one can imagine, the analog signals produced by a transducer are typically weak, in other words, the electrical amplitude is small. To alleviate this, and prepare the signal to be passed to the subsequent processing block, an electrical amplifier is typically used. Therefore, in order to use the electrical signal, amplification is of utmost importance, which highlights the fundamental nature of the electrical amplifiers.

It should be noted that amplifiers are not only needed in the early stages of electrical manipulation, but also in many of the latter blocks. This is evident in data converters, voltage-controlled oscillators and bandgap references.

1.3 The Ideal Op Amp

Electrical augmentation can come in many shapes and sizes, but what if one could imagine a perfect (fictitious) amplifier, what would it look like? Well, it would be convenient if the gain (ratio of output signal to input signal) was very large, or even infinite. At this point some may jump up in protest underlining the fact that electrical signals are finite[2], but this

[2] Any physical gain must be finite to take place in reality.

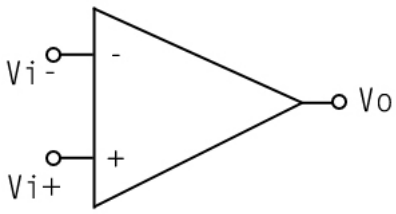

Figure 1.1: The circuit symbol for an ideal op amp.

can be considered later. Another convenient feature would be for the amplification to be linear, meaning that all electrical levels are handled equally no matter the amplitude. The analog signals will also vary with frequency, so an amplifier that can treat all frequencies the same, in other words, having an infinite bandwidth, would be attractive. Lastly, random noise can be a nuisance, so an amplifier that can handle differential signals would alleviate problems. To recapitulate, the specifications for the fictitious amplifier are:

- Infinite gain
- Linear gain
- Infinite bandwidth
- Differential signal capabilities

These describe an ideal operational amplifier, a key piece to any electronic circuit, and is typically depicted by the diagram in figure 1.1.

Op amps can have differential outputs to improve noise reduction as well as increased signal swing; however, for simplicity the single-ended implementation will be discussed. The mathematical equation that governs the op amp is:

$$V_o = A(V_{i+} - V_{i-}) \tag{1.1}$$

Where V_o is the output signal, A is the gain, V_{i+} is the positive input signal and V_{i-} is the negative input signal. Note, these discussions are being done with voltages, but the signals could also be currents. To get an idea of what the function A is actually doing, consider figure 1.2.

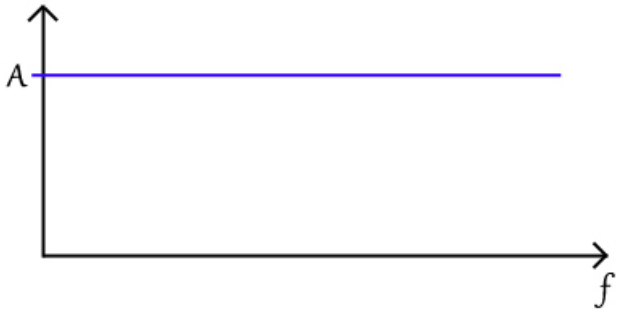

Figure 1.2: A plot of gain (in decibels) versus frequency for an ideal op amp.

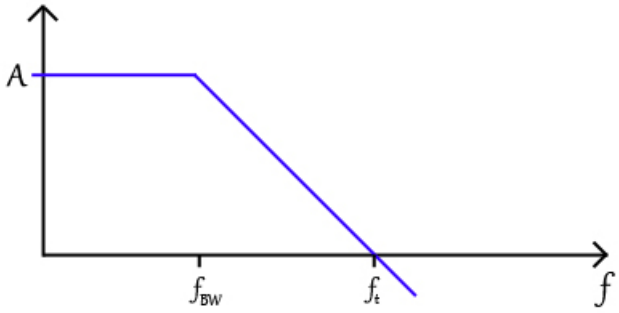

Figure 1.3: A plot of gain (in decibels) versus frequency (log-scale) for a typical amplifier, showing a finite gain and low-pass characteristics.

It is clear that the transfer function A plays a critical role in the op amp behaviour, and with infinite bandwidth and gain there are no real problems. Unfortunately, it is time to step out of the dream and face reality. The transfer function usually takes the shape shown in figure 1.3, where the gain eventually falls off due to physical limitations.

The plot shows that there is now a finite bandwidth and gain, which means that the analog signals are now confined to a finite output and frequencies are no longer all treated the same. This highlights limitations that come about due to the physical nature of the op amp; however, since the ideal amplifier cannot be readily built, this creates competition (and jobs) for those who can create the "best" op amp.

In order for those competitors (herein called designers) to communicate, it is necessary to define some terms depicted in figure 1.3. First of all, the value of the gain at low frequencies is typically called the DC gain or A_o. The point at which the gain drops by 3

decibels, or the power drops by half, is called the 3 dB bandwidth or ω_{3dB}. The bandwidth frequency divides the plot between a constant gain, and a decreasing gain that eventually drops below 0 decibels (the decrease in gain is at a rate of 20 dB/decade). Ostensibly, the unity-gain frequency is the point when the gain is equal to unity of 0 dB. These three terms, DC gain, bandwidth and unity-gain frequency can all be related [1]. The plot shown in figure 1.3 is described by the following complex equation, many will recognize this as a low-pass filter - which it is:

$$A(j\omega) = \frac{A_0}{1 + j\omega/\omega_{3dB}} \tag{1.2}$$

Using the awareness of unity-gain frequency, ω_t, to find a useful relationship...

$$|A(j\omega_t| = \left|\frac{A_0}{1 + j\omega_t/\omega_{3dB}}\right| = 1 \tag{1.3}$$

$$|A_0| = |(1 + j\omega_t/\omega_{3dB})| \tag{1.4}$$

$$|A_0| = \sqrt{(1^2 + \omega_t^2/\omega_{3dB}^2)} \approx \frac{\omega_t}{\omega_{3db}} \tag{1.5}$$

The approximation can be made since $\omega_t > \omega_{3db}$. This leads to an important relationship that relates gain, bandwidth and unity-gain frequency.

$$\omega_t = A_0\omega_{3db} \tag{1.6}$$

Due to equation 1.6, the unity-gain frequency is often used interchangeably with the gain-bandwidth product (GBW). In one-pole low-pass systems, the unity-gain and gain-bandwidth product are equal. However, it is important for the reader to notice that this is not always the case; consider the plot in figure 1.4.

Here, the unity-gain frequency and the gain-bandwidth product are not equal, and give insight to different characteristics of the relationship. Typically, the GBW is the maximum achievable unity-gain frequency, or in other words, it often overestimates the unity-gain frequency. The occurrence and location of the higher poles will be discussed in the next chapter, but before going into the inner workings of the op amp, it is important to review some common circuit configurations.

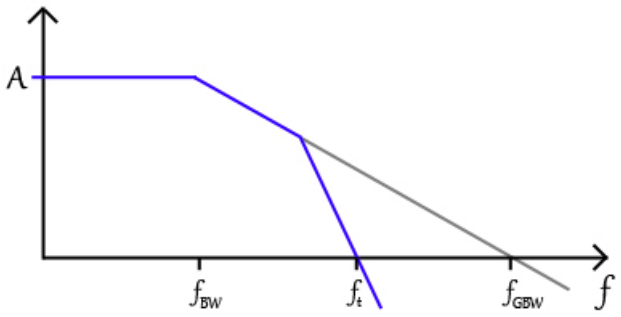

Figure 1.4: Two pole system, plotting the gain as a function of frequency. Here the unity-gain frequency and the gain-bandwidth product are not equal.

1.4 Op Amp Configurations

After discussing that there will be limitations to the ideal amplifier, it is time to begin thinking about how one would actually use this device. Many of the concepts presented here are based on the discussions in Chapters I, II and X of Sedra and Smith (sixth edition) [1].

Imagine that an audio amplifier is needed. The audio band is naturally between 20 Hz - 20 kHz and the transducer produces signals in the range of 10-100 μV. However, the signals are too low and hence too sensitive to be processed by the proceeding circuity, so augmentation is needed. The data converters that take the analog signal and digitize it for better processing would prefer signals in the range of 1-10 mV. Hence, an amplification factor (gain) of 100 V/V is required, with a bandwidth larger than 20 kHz. An amplifier is designed with a gain of 100 V/V and a bandwidth of 40 kHz. However, after being manufactured, the chip returns and has a gain of 78 V/V and a bandwidth of 62 kHz. This significant change in performance comes from physical deviations at the fabrication level, which causes problems for designers. Luckily, a clever solution has been employed to get around this problem, and includes better control over the gain using feedback [2].

There are two ways to connect the op amp in feedback, one is to connect the output to the positive terminal, and the other is to connect the output to the negative terminal. Positive feedback may lead to instability, so it is advisable to make the connection to the negative terminal as shown below in figure 1.5.

Examining this circuit with the pre-defined equations:

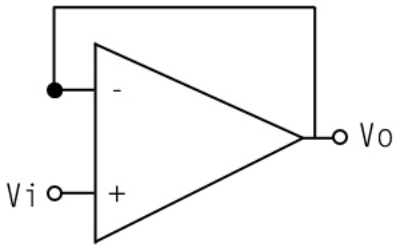

Figure 1.5: Unity-gain op amp configuration.

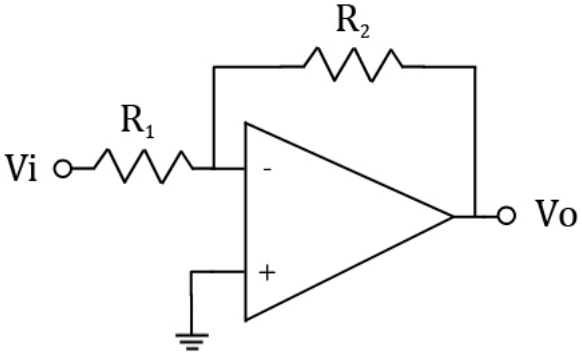

Figure 1.6: An inverting op amp topology, where R1 and R2 compose the feedback network.

$$V_o = A(V_+ - V_-) \tag{1.7}$$

$$V_o = \frac{V_+}{1 + 1/A} \approx V_+ \tag{1.8}$$

The approximation can be made if A is much larger than unity. At this point, the impact of feedback may seem irrelevant; it is clear that this changes the transfer function, but how does this help in design? The key is in the following statement: the transfer function no longer depends on A as long as $1/A < 1$, in other words, as long as A is large. This makes designing simple, and more reliable since there is no longer a specific target for the op amp, just as long as the gain is large (typically 1000 V/V to 100000 V/V).

One can now begin to explore more feedback circuits, like the "inverting op amp" in figure 1.6. To obtain insight concerning its behaviour, the same analysis will be done.

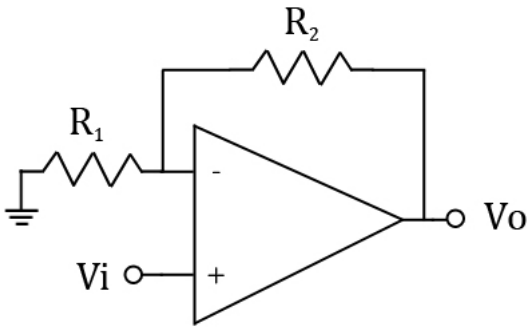

Figure 1.7: Non-inverting op amp topology.

$$\frac{V_i + V_o/A}{R_1} = \frac{-V_o/A - V_o}{R_2} \tag{1.9}$$

$$\frac{V_o}{V_i} = \frac{-R_2/R_1}{1 + \frac{1+R_2/R_1}{A}} \tag{1.10}$$

If the designer does their job and ensures that $A \gg 1 + R_2/R_1$, then the gain is solely based on the *passive* feedback components R_1 and R_2.

The non-inverting op amp, shown in figure 1.7, is the answer to those who do not want a minus sign in the transfer function. To do this, one applies the input to the op amp's non-inverting terminal, while keeping feedback circuitry connected to the negative input.

$$\frac{V_o}{V_i} = \frac{1 + R_2/R_1}{1 + \frac{1+R_2/R_1}{A}} \tag{1.11}$$

A similar relationship is obtained for the gain in equation 1.11. Shifting the gain dependency from A to the passive resistors is advantageous since passive components, like resistors and capacitors, are more easily controlled at the fabrication level.[3]

Behaviour is more complex when capacitors are present in the feedback network. Since capacitors are frequency dependant elements, the feedback network becomes frequency dependant which leads to op amps that can perform frequency-varying tasks like integration and derivation.

[3]To be a little more specific, capacitor *ratios* can be precisely predicted for most fabrication processes.

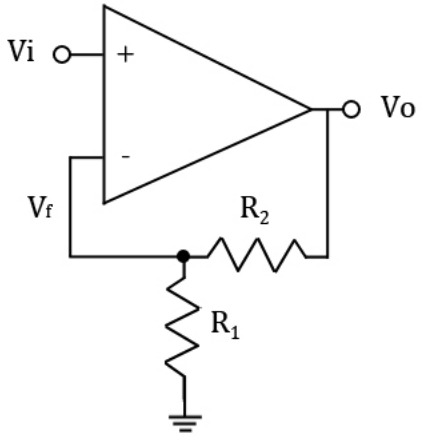

Figure 1.8: The non-inverting op amp, redrawn to emphasize the feedback network.

However, the usefulness of understanding integrator or differentiator circuits is not critical for this work and will be left to the experts. Before going further, it is useful to look into the feedback network in a little more detail; specifically, the stability conditions.

1.5 Feedback and Stability

The feedback network composed of R_1 and R_2 is taking some of the output signal and feeding it back to the input. This process is better highlighted in the non-inverting op amp, see figure 1.8.

The passive resistor network forms a voltage divider of the output voltage, meaning that the amount of output signal can be anywhere from a factor of 0 to 1 depending on the relative values of R_1 and R_2. At this point, one can define another parameter called the "feedback factor", which is a quantification of how much output signal gets fed back to the input. Note that the gain is shown in the following analysis.

$$\frac{V_o}{V_f} = \frac{1}{\beta} = \frac{R_1 + R_2}{R_1} = 1 + R_2/R_1 \tag{1.12}$$

This result should look familiar: it is the gain that was derived previously. Rewriting equation 1.11:

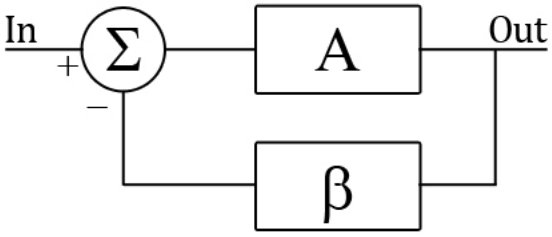

Figure 1.9: The traditional feedback block diagram, separating the gain and feedback network.

$$\frac{V_o}{V_i} = \frac{1/\beta}{1 + 1/A\beta} = \frac{A}{1 + A\beta} \tag{1.13}$$

The obtained equation is critical when speaking about feedback and stability. However, it is typically derived from the more general block diagram shown in figure 1.9.

By analyzing this figure, it is straightforward to arrive at the general negative feedback equation:

$$(In - \beta Out)A = Out \tag{1.14}$$

$$\frac{Out}{In} = \frac{A}{1 + A\beta} \tag{1.15}$$

This equation is typically referred to as the 'closed-loop' gain equation or A_{CL}, since it highlights the feedback loop which closes the system. As discussed, such feedback offers many advantages like better gain control, increased linearity and noise reduction (from active components inside the op amp); however, feedback also introduces additional challenges, mainly relating to the stability of the circuit. Positive feedback was avoided since this could produce an output signal even if no input was applied. If both A and B blocks remained constant, then the proposed circuit would always be stable; unfortunately, this is not the case. It has already been highlighted that A is actually a function of frequency, and β can also be whether or not the designer is aware (parasitics). Returning to the transfer function of A:

$$A(s) = \frac{A_0}{1 + j\omega/\omega_{3dB}} = \frac{A_0}{\sqrt{1 + \omega^2/\omega_{3dB}^2}} e^{j\phi} \tag{1.16}$$

By rewriting the transfer function in polar form, a subtle yet extremely important realization is made: not only does the gain change with frequency but also the signal's phase (denoted as ϕ). This is important since the only expected phase change has been the 180^o shift due to the inverting input. While considering the transfer function alone, this may not seem of interest, but when remembering that the amplifier block is placed in a feedback system it means that, at some frequency, there may be an unpredicted 180^o phase shift. It is imperative that the total phase shift around the loop consisting of A and β is not 180^o, since this will result in the feedback signal adding to the input instead of subtracting. Specifically, this would not be a problem if the loop gain is also below unity, since the signal will die out. So, one can summarize the (Nyquist) stability criterion as follows [3]:

$$A(j\omega_{180})\beta(j\omega_{180}) < 1 \tag{1.17}$$

To understand the applicability of said criterion, consider the Bode plot which has been littered with poles in figure 1.10.

The "danger zone" is defined as all frequencies where the phase shift is 180^o and the gain is larger than unity - this is not where a designer would like to be unless oscillation is intentional. Hence, to ensure an op amp design is stable, a parameter is defined which quantitatively illustrates the relationship between gain and frequency (at the point of instability): phase margin (PM). The safety factor known as phase margin is best defined by using the plot above. With an understanding of the Bode plot, the phase margin is defined by equation 1.18 below:

$$PM = 180^o - \sum_i \tan^{-1}\left(\frac{\omega_t}{\omega_i}\right) \tag{1.18}$$

Where ω_i is the system's pole frequencies. It is also important to note that for multiple pole systems, a PM of 0^o may not lead to a completely stable circuit. Often, with phase margin values below 45^o, the output experiences significant oscillations which significantly affects settling behaviour. For design, a phase margin of at least 65^o is typically required to yield suitable transient settling. This number may seem arbitrary, but it tends to work well so designers continue to use it.

In most circuits, the phase margin is not suitable initially. However, since the loop gain is composed of two terms, A and β, engineers will take advantage of the ability to control A in order to hit the PM specification while β is usually set by the desired closed-loop gain. Furthermore, some designers go as far as tuning the pole locations. The process of

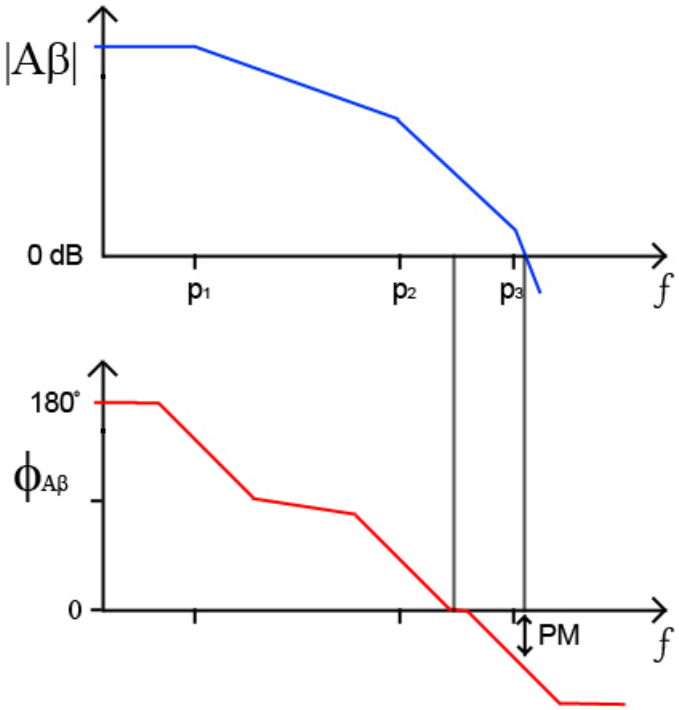

Figure 1.10: A Bode plot with multiple poles to highlight the stability concerns.

changing the transfer function to hit stability specification is called compensation - one of the primary topics of today's op amp design.

Speaking of op amp design, it is now time to dive into today's formulations of the guts that actually make amplifiers. Note that this introduction has not been written to give a complete picture of each topic, but to set the stage for the rest of this work. Relevant topics will be discussed in the following chapters, while the specific problem is introduced in the next section.

Chapter 2

Literature Review

In this chapter, fundamentals of op amp design will be discussed; beginning with a quick introduction to op amp topologies and current design strategies. Previously, op amps were discussed as a black box, given terminal characteristics that were deemed suitable for the envisioned application. However, this black box must be physically built. Modern electronics use a fundamental component, called a transistor, to build op amps. Transistors, in general, are 3-terminal active devices that allow the control of one terminal's behaviour by adjusting the other two. This separation, or independence, is what allows amplifiers to work.

Through the 1960s into the early 1990s, the transistor was implemented by using a p-n-p (or n-p-n) semiconductor junction to create the device known as a bipolar junction transistor (BJT) [4]. BJT type op amps typically led to very high gain performance and were reliable. However, around the early 1990s, digital electronics began to dominate the chip-space and these circuits required a different transistor implementation called a metal-oxide-semiconductor field-effect transistor (MOSFET or simply MOS) [5]. Similarly to the doping variations in BJTs, MOS transistors can be implemented using a conductive n-doped or p-doped channel, leading to NMOS and PMOS devices. MOSFETs were advantageous since their leakage currents were low, which opened the door for solid-sate memory as well as lower power consumption (mobile) applications. Yet, MOSFETs typically show much lower gain performance for congruent geometries [6].

This work will focus on MOSFET implementations, by reason of most state-of-the-art electronic processes are based in MOS technologies. Knowing this, it is now time to review some op amp designs and discuss their performance.

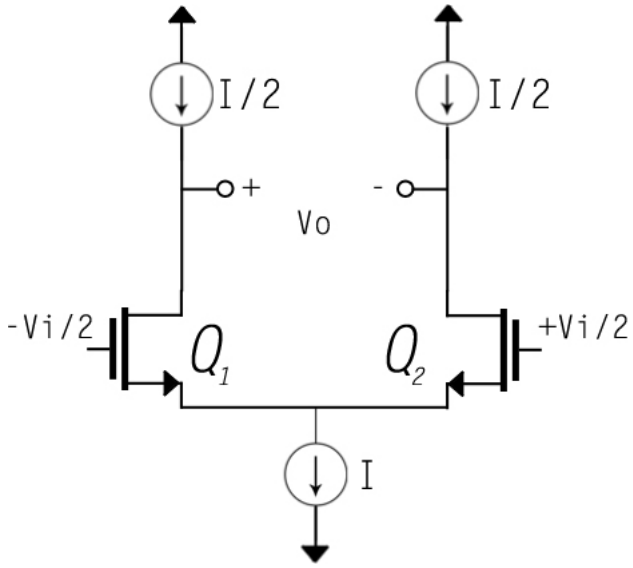

Figure 2.1: Differential pair (NMOS) topology where biasing is done with ideal current sources.

2.1 Types of Op Amps

The first challenge to address is how one might build a differential input amplifier. Thankfully, this problem has been solved through the differential pair circuit - introduced back in the days of the vacuum tube [7]. It is a core component to all op amp circuits and is presented in figure 2.1.

Initially, the functionality of this circuit may not be apparent, but when looking at the small signal model it is clear that there is a push-pull current control effect. Using the small-signal model (note that this is a simplified perspective for clarity reasons), it is possible to derive the transfer function:

$$V_{o+} - V_{o-} = g_m r_o [V_{i+} - V_{i-}] \tag{2.1}$$

Where V_{o+} and V_{o-} are outputs voltages, V_{i+} and V_{i-} are the input voltages, g_m is the transconductance and r_o is the finite output resistance of the transistor. Equation 2.1 is already close to the relationship of the ideal op amp discussed in the introduction. The gain, A, is given by the transconductance (g_m) and the finite output resistance (r_o). A key difference is that the output is differential, which is sometimes necessary, but in this work

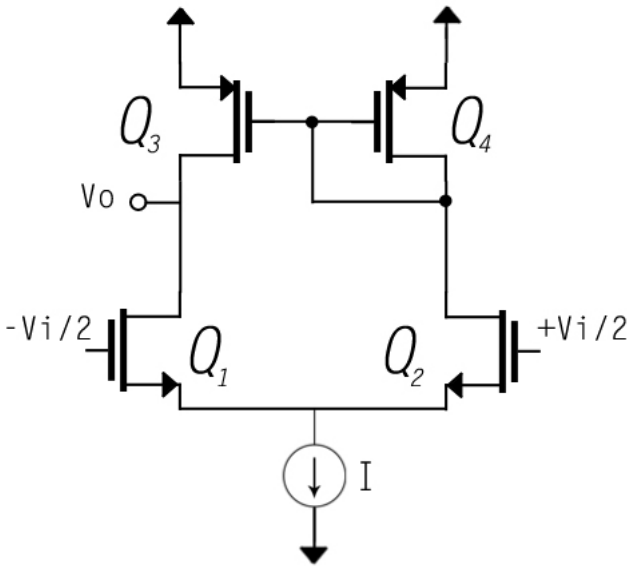

Figure 2.2: NMOS differential pair circuit with PMOS current-mirror active load.

a single-ended output will suffice (avoiding the additional complexities of common-mode feedback). To obtain a single-ended output, the following modification can be done: replace the current source loads with MOS transistors, and connecting V_{o-} to a diode connected device. Consider the circuit in figure 2.2, and it can be shown that:

$$V_o = \frac{g_m r_o}{2}[V_{i+} - V_{i-}] \tag{2.2}$$

Where it has been assumed that all devices are matched ($r_{on} = r_{op}$). At this point, the differential pair with an active load yields a transfer function exactly the same as the fictional op amp where $A = g_m r_o/2$. Note that the gain has decreased by a factor of 2 since the current devices are no longer ideal.

Often times, a gain of $g_m r_o/2$ is not enough. For example, in a 180nm process, g_m is in the range of 6 mA/V while r_o is in the range of 7.5 $k\Omega$. One can change these values to increase the gain, but at the trade-off of some other specifications.

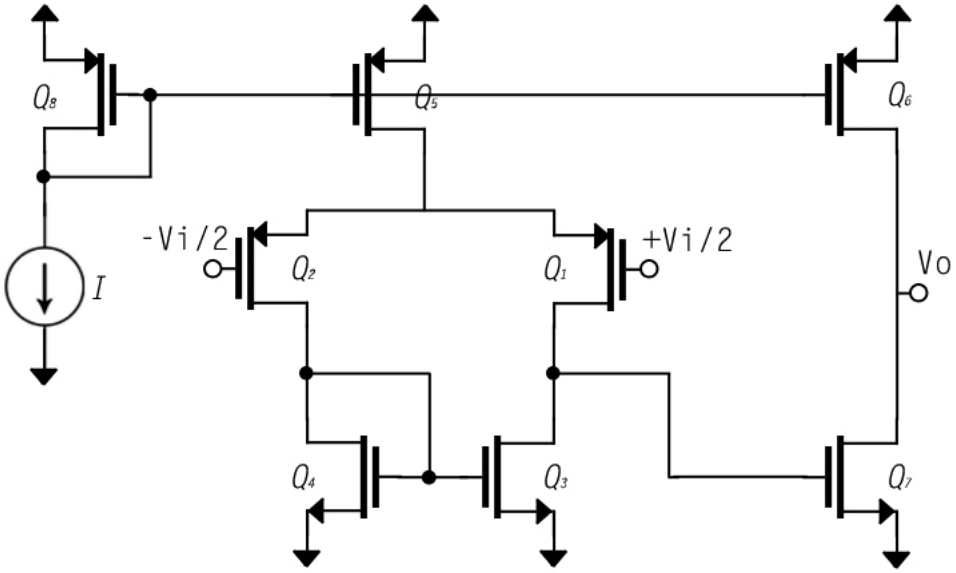

Figure 2.3: A two-stage CMOS op amp, with a PMOS input differential pair and NMOS common-source second stage.

2.2 Two-Stage CMOS Op Amp

The most popular way to deal with low-gain amplifiers is the concept of stages, where additional amplifiers are added after the differential pair to boost the gain. An example of such an implementation is shown in figure 2.3, and is called a two-stage complementary metal-oxide-semiconductor (CMOS) op amp.

Here, a common-source amplifier is tagged-on after the differential pair in hopes to boost the gain. Theoretically, and by using the gain derived previously, the output voltage is given by equation 2.3.

$$V_o = (g_{m1}R_1)(g_{m7}R_2)[V_{i+} - V_{i-}] \tag{2.3}$$

Where R_1 and R_2 are the output resistances of the first stage and second stage, respectively. Although the two-stage amplifier does increase gain, it introduces an additional pole in the system, which causes concern for stability. Previously, the differential pair contained one pole, leading to a worst case phase margin of 90^o. However, the two-stage system has two poles, which results in a worst case phase margin of 0^o – well below the required 65^o.

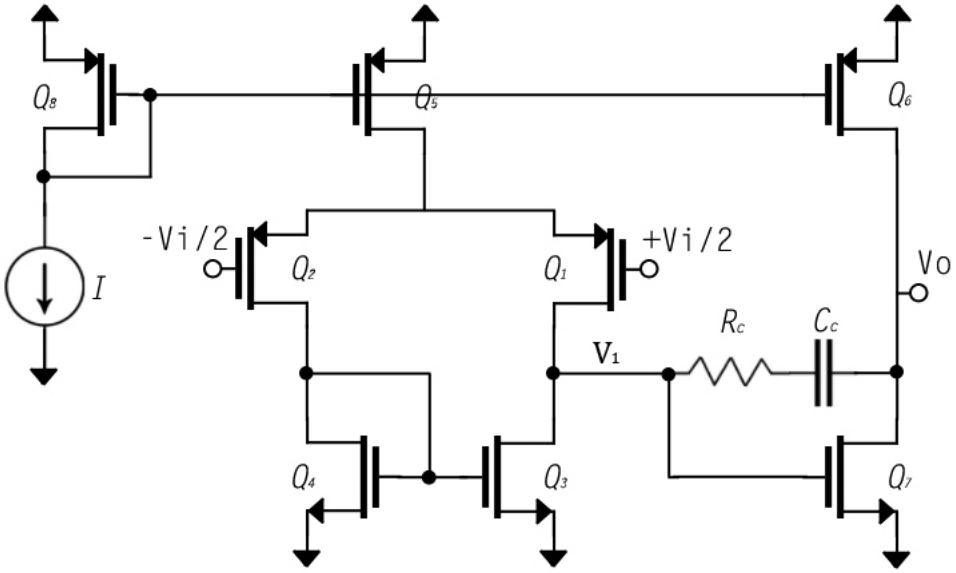

Figure 2.4: Two-stage CMOS op amp with Cc and Rc included.

AC analysis can be used to determine the frequency response of the two-stage amplifier, if the gate-to-drain capacitances are neglected, the resulting transfer function is shown in equation 2.4.

$$A(s) = \frac{(g_{m1}R_1)(g_{m7}R_2)}{(1 + sR_1C_1)(1 + sR_2C_2)} \tag{2.4}$$

Now, there are two poles that can lead to problems in the phase margin; furthermore, the pole locations are difficult to control since they share parameters with the DC gain. A clever solution to this problem has already been established, luckily, and involves adding a compensation capacitor, C_C, and a series resistor, R_C, between the two stages, as shown in figure 2.4.

The addition of the coupling capacitor adds complexity to the analysis, since the two-stages can now interact. Ignoring R_C leads to the following two equations (by performing Kirchoff's current law at V_1 and V_o nodes), which can be used to determine the overall transfer function.

$$0 = i_i + \frac{V_1}{R_1} + sC_1V_1 - i_C \tag{2.5}$$

$$0 = g_{m7} + \frac{V_o}{R_2} + V_o s C_2 + (V_o - V_1) s C_C \tag{2.6}$$

Using these equations, it is possible (but not trivial) to derive the following transfer function:

$$\frac{V_o}{V_i} = \frac{g_{m1} g_{m2} R_1 R_2 (1 - s C_C / g_m)}{1 + s[\alpha] + s^2[\beta]} \tag{2.7}$$

$$\alpha = R_2 C_2 + R_2 C_C + R_1 C_1 + R_1 C_C + R_1 R_2 g_{m7} C_C \tag{2.8}$$

$$\beta = R_1 R_2 C_C (C_2 + C_C) + R_1 R_2 C_1 (C_2 + C_C) - R_1 R_2 C_C^2 \tag{2.9}$$

It is clear that the transfer function contains two poles and one zero, but to properly extract the pole locations, factorization is needed.

$$(1 + \frac{s}{\omega_1})(1 + \frac{s}{\omega_2}) = 1 + s/\omega_1 + s/\omega_2 + s^2/(\omega_1 \omega_2) \tag{2.10}$$

Assuming the poles are widely separated...

$$(1 + \frac{s}{\omega_1})(1 + \frac{s}{\omega_2}) \approx 1 + s/\omega_1 + s^2/(\omega_1 \omega_2) \tag{2.11}$$

Implying,

$$\omega_1 = \frac{1}{\alpha} \tag{2.12}$$

Now, since $R_{1/2} g_{m7}$ is typically much larger than unity, then α reduces to one term and yields the first (dominant) pole location, shown in equation 2.13.

$$\omega_1 = \frac{1}{R_1 R_2 g_{m7} C_C} \tag{2.13}$$

The dominant pole location can be used to find the second pole, shown in equation 2.15.

$$\omega_1 = \frac{R_1 R_2 g_{m7} C_C}{R_1 R_2 C_C (C_2 + C_C) + R_1 R_2 C_1 (C_2 + C_C) - R_1 R_2 C_C^2} \tag{2.14}$$

$$\omega_2 = \frac{g_{m7} C_C}{C_C C_2 + C_1 C_2 + C_1 C_C} \tag{2.15}$$

The two pole locations are now defined, and can be used in design. At this point, it is important to note the affect of C_C on both pole locations (an increase in C_C causes the poles to split apart). And finally, a zero has been introduced. This zero causes problems since it accelerates the decrease in phase, which will have adverse effects on the phase. Designers can obtain better control over the zero location by placing a resistor in series with the coupling capacitor, as shown in figure 2.4, denoted as R_C [8]. The zero occurs when the current "leaking through the coupling capacitor (or feeding-through) equals the current being sunk by the current-controlled source, resulting in no current flowing through C_2 and R_2, meaning that $V_o = 0$. Without R_C, this occurs at:

$$(V_1 - V_o) s C_C = g_{m7} V_1 \tag{2.16}$$

$$s = j\omega = g_{m7}/C_C \tag{2.17}$$

With R_C introduced:

$$\frac{V_1 - V_o}{R_C + 1/sC_C} = g_{m7} V_1 \tag{2.18}$$

$$\omega_z = \frac{g_{m7}}{C_C (1 - g_{m7} R_C)} \tag{2.19}$$

Now, it is possible to control the zero location by tuning R_C, as shown in equation 2.19. Controlling the value of R_C and C_C becomes the principal way to compensate the two-stage op amp. Specifically, by increasing C_C, one can split the two-poles to give a suitable phase margin. Furthermore, it is possible to design R_C to now add to the phase, yielding a better phase margin. The discussion of compensation-driven design will be revisited later in this chapter.

Returning to the transfer function shown in 2.20, where ω_z, ω_1 and ω_2 are given by 2.19, 2.13 and 2.15, respectively; it is now possible to extract useful DC and AC information.

$$\frac{V_o}{V_i} = \frac{g_{m1}R_1 g_{m2}R_2(1 - s/\omega_z)}{(1 + s/\omega_1)(1 + s/\omega_2)} \tag{2.20}$$

The reader should prepare to view many characterizing equations that may not seem to have any comprehensive relevance; however, all are needed in the design process and will be discussed later on. With this, the DC gain can be readily identifies from 2.20 as:

$$A_0 = g_{m1}R_1 g_{m2}R_2 \tag{2.21}$$

Thus, to control the gain, one can change the size of the transistor as well as the bias current. Frequency response, or the overall speed, is also of interest and can be found by using the equation 2.22 derived in the previous chapter:

$$GBW = A_0\omega_{3dB} = g_{m1}/C_C \tag{2.22}$$

To be complete, there are several other amplifier characteristics that can be defined in order to properly motivate the next discussion. One of these is slew rate, which is defined as the maximum rate of change for the amplifier output, and can be shown in equation 2.23 [9].

$$SR = I/C_C \tag{2.23}$$

This particular equation is due to the structure of the differential pair, and can change for different amplifier formations. Another very important performance metric is called the systematic offset voltage. In this work, the offset voltage is defined as the DC voltage that must be applied to the positive input in order for the amplifier's output to be exactly halfway between the rails.

Many other characteristics like output swing, common-mode rejection ratio and power dissipation can also be defined, but are not of utmost importance for this work.

At this point, the reader may be confused, and wonder why all these definitions are needed. Op amps are not perfect, so specifications are needed to communicate to the user what advantages certain op amps have over others. Additionally, such definitions aid in the design process from an optimization standpoint.

2.3 Design Methods in Literature

As alluded to previously, most design strategies rely heavily on the discussion on compensation, leaving specifications up to the designer. After examining several textbooks and papers, most point to a publication by Gray and Meyer entitled "MOS Operational Amplifier Design – A tutorial overview" [10]. In this paper, the authors highlight the design metrics of the two-stage op amp, and derive the pertaining equations that govern them. The authors' summarize the specifications of an amplifier in the way shown in table 2.1. Note that the power-supply rejection ratio and common-mode rejection ratio are denoted as PSRR and CMRR, respectively.

Table 2.1: The typical op amp specifications highlighted by Gray and Meyer, 4um CMOS

DC gain	5000
Settling time, 1V Step, C_L=5pF	500 ns
Equiv. input noise, 1 KHz	100 nV/$\sqrt{Hz}$
PSRR, DC	90 dB
PSRR, 1 KHz	60 dB
PSRR, 50 kHz	40 dB
Supply capacitance	1 fF
Power dissipation	0.5 mW
Unity-gain frequency	4 MHz
Die area	75 mm^2
Systematic offset	0.1 mV
Random offset std. dev.	2 mV
CMRR	80 dB
CM Range	within 1V of supply

The $4/mum$ process was common in the 1980s, so many of the specifications shown above are quite dated. Nonetheless, it gives a good idea on just how many variables a designer must keep in mind. In addition, the optimization of different specifications depends greatly on the intended application, so having an all encompassing design strategy is difficult. Instead, authors typically outline "rules of thumb" that can give helpful tips on designing – being fairly independent of the application. A point form summary of these suggestions, by Grey and Meyer, are shown below (they appear in no particular order, but are numbered for ease of referencing later on):

1. For a constant current, the gain decreases for decreasing length or width - this means

that for high gain applications, the device size is somewhat fixed

2. For constant device size, the gain increases as the current decreases (up to the point of sub-threshold)

3. Systematic offset voltage can be avoided by obeying the following equation [1]:

$$\frac{(W/L)_4}{(W/L)_7} = \frac{(W/L)_3}{(W/L)_7} = \left(\frac{1}{2}\right)\frac{(W/L)_6}{(W/L)_5} \tag{2.24}$$

4. The zero vanishes when R_C is made equal to $1/g_{m7}$. The resistor can be further increased so that the zero is place in the left half-plane to improve the amplifier phase margin

5. The use of load capacitances of the same order as the compensation capacitor will tend to degrade the unity-gain phase margin, due to the encroachment of the non-dominant pole

The list above is by no means a complete summary of the ref. [10], but outlines important topics that are relevant for this work. It is time to investigate the points, beginning with item 1 where device size is said to have a strong affect on gain. Starting with equation 2.21, using the first-stage component and substituting in the relevant values leads to:

$$A_{01} = \sqrt{\frac{1}{2}\mu_p C_{ox}\left(\frac{W}{L}\right) I_{D1}}\left(\frac{1}{\lambda I_{D1}}\right) \tag{2.25}$$

Where μ_p is the PMOS mobility, C_{ox} is the oxide capacitance and λ is the output impedance constant. It is important to note that λ contains a length component, meaning that the equation can be simplified further. Defining $\lambda' = \lambda L$, where λ' is a constant for a given technology:

$$A_{01} = \sqrt{\frac{1}{2}\mu_p C_{ox}\left(\frac{W}{L}\right) I_{D1}}\left(\frac{L}{\lambda' I_{D1}}\right) \tag{2.26}$$

$$A_{01} = \gamma\sqrt{\left(\frac{WL}{I_{D1}}\right)} \tag{2.27}$$

[1] Devices are numbered with respect to figure 2.4.

Where γ contains all the constant parameters. From equation 2.27, the two first points are rather obvious. It is clear that the gain increases with device size (width and length) while the gain decreases with increased current. An analogous equation can be derived for the second stage; hence, equation 2.27 offers a key insight into design. A slight modification to the equation can be done, which will have a huge impact; consider equation 2.28.

$$A_{01} = \gamma\sqrt{\left(\frac{W}{I_{D1}}\right)L} \tag{2.28}$$

The three variables have been written in a way that only has two aspects to change: the current density defined as I_{D1}/W and the length L. To make it clear why this distinction has been made, consider the equation derived for the second pole location in equation 2.15. This second pole ultimately determines the unity-gain frequency due to phase margin. [2] A similar analysis can be done by assuming that C_C is much larger than the parasitic capacitances C_1 and C_2.

$$\omega_2 \approx \frac{g_{m7}}{C_1 + C_2} \tag{2.29}$$

Now, C_2 typically contains the load capacitor, but it is of more interest to understand how the transistors dictate speed rather than the application. Thus, C_2 only contains the parasitic capacitances.

$$\omega_2 \approx \frac{g_{m7}}{(C_{db2} + C_{db4} + C_{gs7}) + (C_{db7} + C_{db6})} \tag{2.30}$$

$$\omega_2 \approx \frac{g_{m7}}{WC} \tag{2.31}$$

$$\omega_2 \approx \frac{\sqrt{\frac{1}{2}\mu_n C_{ox}\left(\frac{W}{L}\right)I_{D2}}}{WC} \tag{2.32}$$

$$\omega_2 \approx \gamma_2\sqrt{\frac{I_{D2}}{W}\frac{1}{L}} \tag{2.33}$$

[2] There must be a fixed ratio between the two as will be discussed later.

This simplification is possible since all parasitics scale with the width of the device; hence, C (capacitance per unit width) contains all capacitive constants.[3] The final result, shown in equation 2.33, highlights the trade-off between gain and speed. To increase the second pole frequency, the current density must be increased or the length of the device must be decreased. This is exactly opposite than the case for gain. Therefore, there is a fundamental trade-off between gain and speed.

The third point made by Gray and Meyer, is that to avoid a systematic offset voltage, equation 2.24 must be obeyed. To understand this, please refer to figure 2.4. Essentially, the NMOS transistors $Q_{3,4,7}$ should all operate with the same current density; under quiescent conditions, this will keep the drain of Q_7 fixed to the drain of Q_4. It should also be noted that reference [10] suggests that the lengths of the NMOS and bias PMOS be equal. For the current densities to be equal:

$$\frac{I_1/2}{W_3} = \frac{I_2}{W_7} \tag{2.34}$$

$$\frac{1}{2}\frac{W_5/W_8}{W_3} = \frac{W_6/W_8}{W_7} \tag{2.35}$$

$$\frac{W_7}{W_3} = 2\frac{W_6}{W_5} \tag{2.36}$$

Unfortunately, ensuring equation 2.36 does not completely eliminate the systematic offset, but it does bring the amplifier close to being balanced. To ensure complete equity, Q_6 should be cascoded since the input PMOS transistors introduce an extra voltage drop.

In regards to the fourth point, it is outlined that the zero can simply be eliminated by setting $R_C = 1/g_{m7}$. This then allows for a effective pole-splitting method of compensation to be used, where the first pole can be lowered while increasing the frequency of the second pole all by controlling the capacitor value. This brings up the final point, which touches on phase margin and how it relates to the compensation capacitor. To discuss this, consider the expression for phase margin below:

$$PM = 180^o - \tan^{-1}\left(\frac{\omega_t}{\omega_{p1}}\right) - \tan^{-1}\left(\frac{\omega_t}{\omega_{p2}}\right) \tag{2.37}$$

[3]Note that all capacitances do not scale with L, and thus, only W can be factored out of the denominator. Also, C_{gs7} is assumed to be much larger than the capacitances from the first stage.

Provided $\omega_{p2} \gg \omega_{p1}$,

$$PM = 90^o - \tan^{-1}\left(\frac{\omega_t}{\omega_{p2}}\right) \tag{2.38}$$

$$\tan(90^o - PM) = k = \frac{\omega_t}{\omega_{p2}} \tag{2.39}$$

From equation 2.22 and 2.15, the intended phase margin can be written in terms of transconductance and dominant capacitance, which are assumed to be C_C and C_L.

$$k = \frac{g_{m1}/C_C}{g_{m7}/C_L} \tag{2.40}$$

For a phase margin of 65^o,

$$k \approx 0.5 = \frac{g_{m1}}{g_{m7}}\frac{C_L}{C_C} \tag{2.41}$$

In contrast to BJTs, MOSFETs typically have similar transconductances, since g_m is proportional to the square-root of current instead of linearly. This implies from equation 2.41 that C_C and C_L should be roughly the same order of magnitude.[4] The concern here is that the load capacitor can be vary large for particular applications, leading to a significant increase in C_C.

Users do not necessarily have a phase margin in mind, they simply want a stable amplifier and for the signal to settle within some amount of time. Specifications do include: DC gain, settling time, input noise, PSRR, CMRR, power dissipation, unity-gain frequency, die area, systematic offset, random offset, and output swing. However, as mentioned previously, the phase margin can significantly impact the amplifier's performance. Due to this, many of the discussions around amplifier design are rooted in compensation and will be discussed in the next section.

2.3.1 Compensation-Driven Design

To address the compensation problem, it is useful to use Carusone, Johns and Martin [11] as an example, whom do not outline a complete design strategy, but do highlight a compensation strategy, as follows:

[4]The capacitors will not be the exact same value, but tend to be relatively similar.

1. Start by choosing $C_C' \approx (\beta g_{m1}/g_{m7})C_L$, where β is the feedback factor
2. Using SPICE, find the frequency, denoted as f', where the phase shift, denoted as ϕ is equal to -125^o. Let the gain at f' be denoted as A', or $A' = A(f')$
3. Let $C_C = C_C'A'$, which will lead to $PM = 55^o$
4. Choose R_C: $R_C = \dfrac{1}{1.7\omega_t C_C}$
5. If not satisfied, increase C_C
6. Replace R_C with a transistor in triode mode

Such a methodology or "rule of thumb" for compensating is typical. Since designers are dealing with two poles, it is possible to outline strategies across various amplifiers. However, using the zero to give phase gains has some adverse effects, as discussed by Kamath [12], most notably that it introduces two time constants which may greatly slow settling time.

Lastly, most of the equations used by literature are based on the square-law model. This model is quite accurate for long devices (>280 nm), but tends to break down with advanced processes. Square-law models are useful in understanding general trade-offs, but exact design is difficult, meaning that, it does get one close to the right answer but not precisely. Modern design makes use of a hybrid approach, meaning that some of the process is based on square-law models, while other results are extracted from computer simulation. Now, all these methodologies are used to try and better predict the final physical product. The best approximations to real chip behaviour are computer models. However, software does not give the designer any insight. So by combining both approaches, one gains both advantages. This type of design is called G_m/I_D (or current density) methodology, and will now be discussed.

2.3.2 Gm/Id Design

To properly understand G_m/I_D design, one must first analyze the fundamental limits of a basic transistor - leading to the performance limits of any circuit [13]. Using figure 2.5, it is possible to quickly write the square-law gain, speed and signal swing.[5] A perfect

[5]Other specifications such as input noise are also possible, but these can usually be controlled by other means, or may not be design limiting.

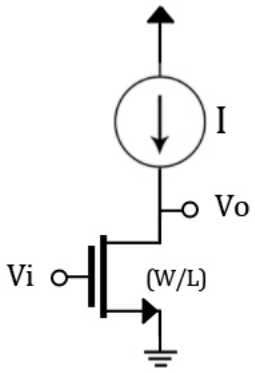

Figure 2.5: A common-source transistor.

common-source transistor is assumed, resulting in all performance being governed by the transistor.

$$A_0 = g_m r_o = \frac{\sqrt{2k'_n}}{\lambda'}\sqrt{\left(\frac{W}{I_D}\right)L} \tag{2.42}$$

$$f_t = \frac{g_m}{C_{gs}+C_{gd}} = \frac{\sqrt{k'_n}}{C'_{gs}+C'_{gd}}\sqrt{\left(\frac{I_D}{W}\right)\frac{1}{L}} \tag{2.43}$$

$$V_{swing} = V_{DD} - V_{ov} = V_{DD} - \frac{2}{k'_n}\left(\frac{I_D}{W}\right)L \tag{2.44}$$

The equations listed above allow one to understand the intrinsic trade-off between gain and speed.[6] However, as technology changes, and designers jump from node to node, the challenge lies in predicting the value of k_n, and other short-channel effects that make modelling difficult.[7] The equations highlight trade-offs but give poor exact predictions. To mitigate this problem, software is used. Setting up the schematic model shown in figure 2.5, the gain, speed and swing can be plotted as a function of current density (I/W) and transistor length (L). The results are shown in figure 2.6.

[6]Emphasis is placed on the fact that this trade-off is not only present in the two-stage op amp, but for the transistor itself, which is the fundamental building block of all modern ICs.

[7]Referred to short-channel effects include drain-induced barrier lowering, velocity saturation and hot electron effect.

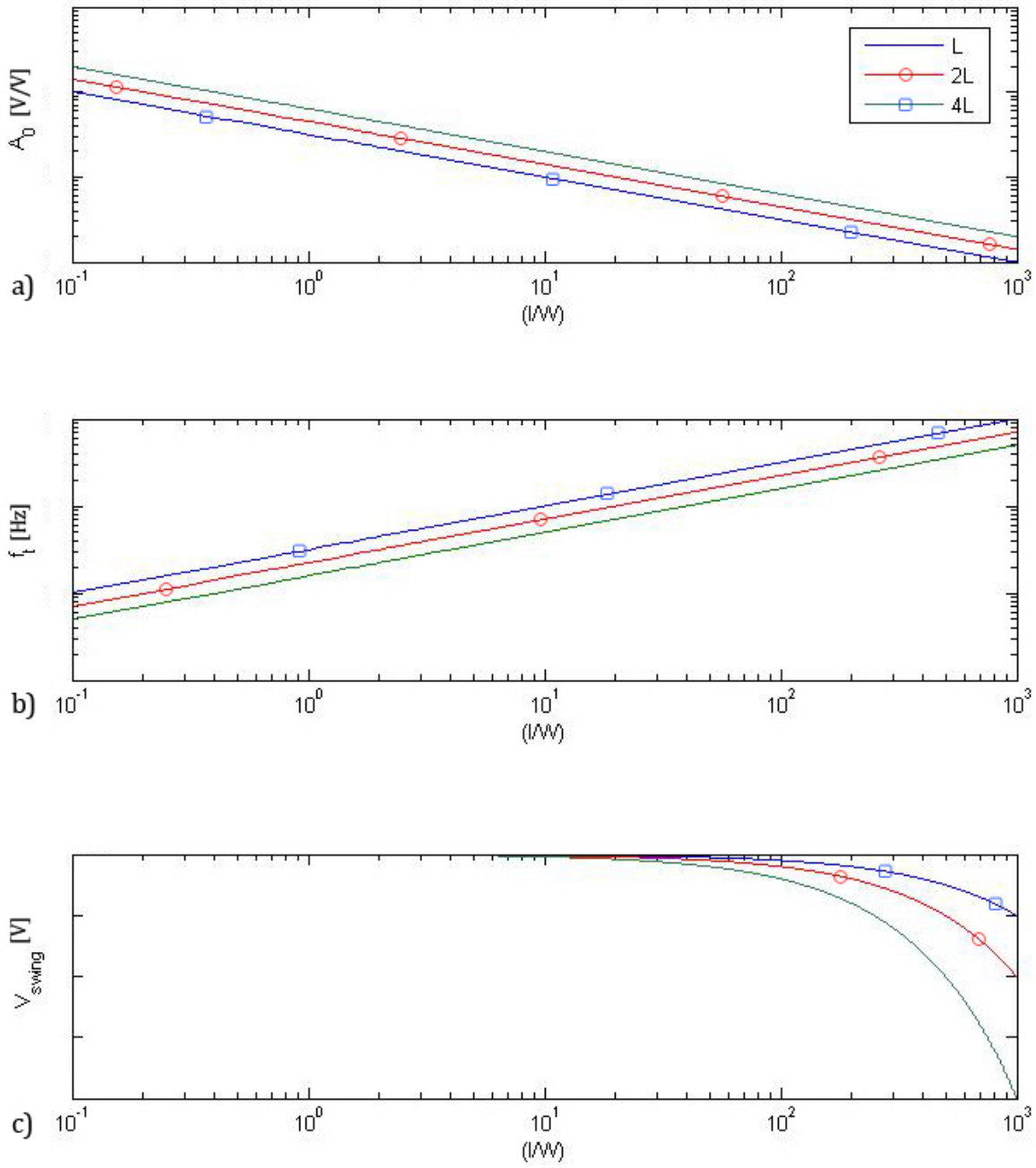

Figure 2.6: A Gm/Id plot, showing the general trends for a) gain, b) speed and c) swing versus current density. Exact values have been omitted to emphasis the relationships with length and current density.

These plots can now be used to estimate the fundamental limits of the transistor along with becoming very powerful, and essential, for design. At this point, it is critical to understand that ideal notions, like threshold voltage, become less relevant at short channel lengths; there is a much larger continuum between triode, saturation and sub-threshold. Luckily, designers can trust the simulation software, provided the devices are accurately reflected, to handle non-ideal effects and only concern themselves with trade-offs of fundamental parameters. [8] Typical design strategies can then have a systematic approach, which evaluates the fundamental characteristics of the transistor (for a particular process).

The design flow is best outlined with an example, but the reader will have to wait for the next chapter if the method is not clear at this point. However, it is useful here to present a design strategy for future reference. According to B. E. Boser, a generic G_m/I_D design flow can be presented [14]:

1. Determine g_m from the design objectives (dynamic range, bandwidth, ...)
2. Pick L
    - Short channel, high f_t
    - Long channel, high gain, good matching
3. Pick g_m/I_D or f_t
    - Large g_m/I_D, low power, larger signal swing
    - Small g_m/I_D, high f_t
4. Determine I_D (from g_m and g_m/I_D)
5. Determine W (from I_D/W, current density chart)

The strategy relies heavily on transconductance determination, which can be a challenge in modern processes. Furthermore, since all applications are very different and present unique specifications, this strategy may not work for all projects. This will be highlighted in the next chapter. However, it does present a general idea of how the design process is executed.

[8] The classic notion of G_m/I_D puts large emphasis on the transconductance. However, transconductance can be difficult to extract from modern processes, so the author of this work will focus on the more fundamental features being width, length and current.

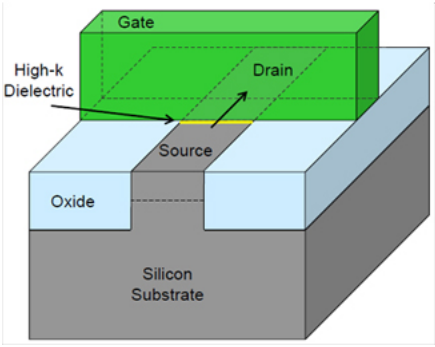

Figure 2.7: The typical planar transistor. Source: Intel

So far, new process nodes have only decreased the actual length of the channel through developments in lithography; however, technology now looks to changing the actual planar architecture of the transistor. Such a change will have a profound impact on analog design and is discussed in the next section.

2.4 Multigate Devices

As channel length decreases, it becomes difficult to properly control the electric field near the channel (inducing wanted behaviour). Secondary effects, mostly edge effects, begin to dominate. This is clear when examining modern planar technologies illustrated in figure 2.7.

Decreasing length (or L) is not without limits. Hence, engineers and scientists have decided to "remove" the channel from the bulk, raise it, and surround it with the gate [15]. Essentially, increasing the gate-channel surface area, while keeping the channel length constant. A diagram of said 3D-transistor is shown in figure 2.8. This type of transistor is also popularly termed a FinFET, since the channel now has a fin type architecture while being surround by the gate.

Having more control of the channel improves many (almost all) transistor characteristics [16]. Furthermore, it allows for smaller transistor lengths, which is the driving force behind technology nodes. From a circuit design perspective, this new transistor structure does not seem to have a huge impact except better (promised) performance. However, there is a key and important change that has occurred, width (or widths in this case) and length are no

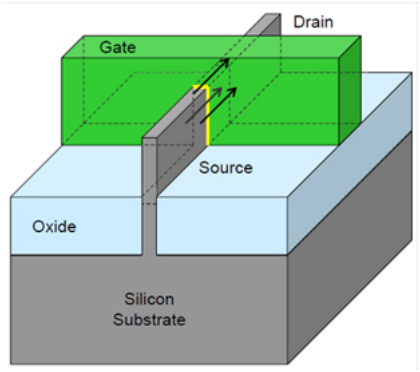

Figure 2.8: A multigate transistor architecture. Source: Intel

longer continuous. This comes down to physical limitations of FinFETs. Such constraints add design complexity and will be discussed in the next section.

2.5 Design Concerns

It is envisioned that, due to physical constraints, only particular lengths and widths will be made available to designers. To most, this change is not critical since the majority of chip technology contains digital circuits, which use only minimum channel length and focus on power consumption; hence, the final result of smaller channel length and increased efficiency (due to the FinFET architecture) is worth the quantization of transistor size for digital designers.

The square-law model does not apply to FinFETs. Luckily, many of the same general (and fundamental) trade-offs are still present. Fortunately, as discussed in the previous section, most designers do not rely heavily on equations, preferring instead simulation plots. A detailed physics analysis of the device is specifically avoided, since the understanding here is to try and take a higher level of abstraction for analog circuits without ignoring key details.

Transistor scaling has presented challenges to analog design. However, some of the concerns mentioned are alleviated by using the G_m/I_D methodology. Yet, losing continuous control over transistor length and width causes problems that G_m/I_D design does not readily solve. In contrast to digital circuitry, which typically always use minimum length devices, analog circuits depend on length to hit gain specifications. The widths of the

devices are also an important parameter that control current density and is typically large compared to length.

For analog circuits, the quantization of width is not troublesome. Most designs use a fixed width, and if more is needed and a second transistor is simply placed in parallel. Such a scheme for increasing the width is effective as well as preferred since it eases layout.

The quantization of length is not so easy to deal with. Although the foundry may offer several length options, there is no guarantee that these options will fit the designers needs. Therefore, one must choose a method for increasing the gain without increasing the transistor length or determine a method to implement custom lengths.

An immediate solution to the fixed-length problem would be cascoding. An example of a cascoded common-source is shown in figure 2.9. Cascoding boosts the gain of an amplifier by a factor of $g_m r_o$ (see Appendix A for more a detailed review of cascodes). As mentioned, the gain of a single transistor is usually not sufficient, so cascoding or multi-staging is used. One could also use a cascode as well as a multi-stage to hit a gain specification while using short transistors. However, cascodes tend to eat up a lot of the signal swing, which is an issue for modern processes that reduce supply voltage. For this reason, the classic two-stage op amp has remained popular even in modern technologies. Cascode architectures also require additional biasing and overhead, which reduces efficiency and increases die area. Lastly, a cascode configuration may boost the gain, but it has worse performance in terms of noise, since noise is determined by the size of the input transistor only, so doubling the length of one transistor would produce less overall noise (this result is derived Appendix B).

From an intuitive perspective, it would be convenient if one could use an approach similar to the one used for transistor width. In other words, it would be easy if increasing transistor length was simply a matter of placing transistors in series (instead of parallel, as is the case for width). Such an idea is the basis for this work, and will be discussed extensively in the next chapter.

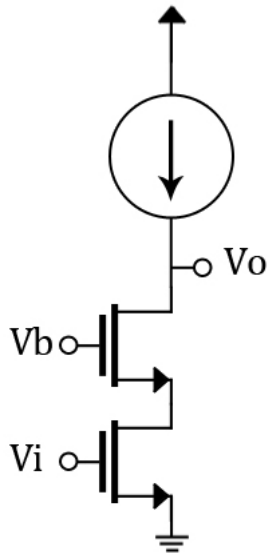

Figure 2.9: A NMOS cascode transistor topology.

Chapter 3

Design

This chapter addresses the fixed-length problem that the FinFET technology presents to analog designers.

3.1 Series-stack

To address the fixed-length problem, the author proposes the following method: to implement a device of twice the length, place two transistors in series. Such an architecture is shown in figure 3.1, and illustrates the differences between the traditional method of simply making the transistor longer (a), and the proposed series method (b).

3.2 Series Stack Derivation

Before moving any further, it is necessary to show that the series design yields the same result as a bulk device. To do this, consider (a) in figure 3.1, the current can be described as (assuming saturation):

$$I_D = \frac{1}{2}\mu_n C_{ox}\left(\frac{W}{2L}\right)(V_{GS} - V_t)^2 \tag{3.1}$$

Taking the derivative of equation 3.1 with respect to V_{GS}, results in the transconductance equation:

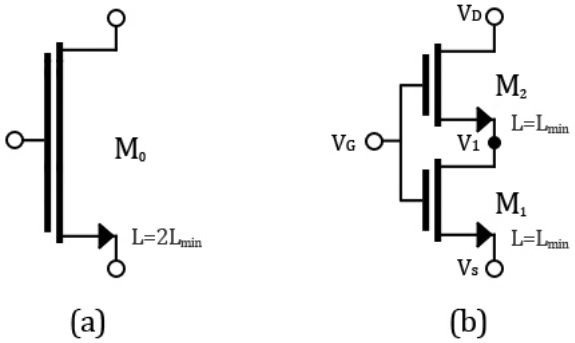

Figure 3.1: Two transistor implementation, (a) being the bulk doubling of the length and (b) being two transistors placed in series.

$$\frac{dI_D}{dV_{GS}} = \mu_n C_{ox} \left(\frac{W}{2L}\right) (V_{GS} - V_t) \tag{3.2}$$

The quantity equation given by 3.2 is typically referred to as the transconductance (g_m).

Moving the attention to (b) of figure 3.1, the bottom transistor can be assumed to be in triode, while the top transistor is assumed to be in saturation. Using this, the current can be described by the two equations below (where V_1 is the middle node):

$$I_{D1} = \mu_n C_{ox} \left(\frac{W}{L}\right) \left[V_{DS1}(V_{GS1} - V_t) - \frac{V_{DS1}^2}{2}\right] \tag{3.3}$$

$$I_{D2} = \frac{1}{2}\mu_n C_{ox} \left(\frac{W}{L}\right) (V_{GS2} - V_t)^2 \tag{3.4}$$

Using some knowledge of the circuit, specifically that $V_{DS1} = V_1 - V_S$ and $V_{GS2} = V_G - V_1 = V_{GS1} - (V_1 - V_S)$ these equations can be simplified,

$$I_{D1} = \mu_n C_{ox} \left(\frac{W}{L}\right) \left[(V_1 - V_S)(V_{GS1} - V_t) - \frac{(V_1 - V_S)^2}{2}\right] \tag{3.5}$$

From eq. 3.4, the following expression can be derived:

$$V_1 - V_S = V_{GS1} - V_t - \sqrt{\frac{2I_{D2}}{\mu_n C_{ox}\left(\frac{W}{L}\right)}} \tag{3.6}$$

Subbing equation 3.6 into equation 3.5:

$$I_{D1} = \mu_n C_{ox}\left(\frac{W}{L}\right)\left[\frac{(V_{GS1} - V_t)^2}{2} - \frac{I_{D2}}{\mu_n C_{ox}\left(\frac{W}{L}\right)}\right] \tag{3.7}$$

$$I_{D1} + I_{D2} = \frac{1}{2}\mu_n C_{ox}\left(\frac{W}{L}\right)(V_{GS1} - V_t)^2 \tag{3.8}$$

From the circuit, it is clear that I_{D1} and I_{D2} are both equal, leading to the final result, shown below in equation 3.9.

$$I_D = \frac{1}{2}\mu_n C_{ox}\left(\frac{W}{2L}\right)(V_{GS1} - V_t)^2 \tag{3.9}$$

By inspecting 3.1 and 3.9, it is evident that the two implementations shown in figure 3.1 are identical. Matching current equations means a matching transconductance, which controls the small-signal characteristics of the transistor. Obviously, there are some simplifying assumptions. None of the non-ideal effects have been taken into account. Furthermore, this is only applicable to DC behaviour, since it is clear that the two circuits will behave differently as a function of frequency due to the extra node in 3.1(b). However, it is not the intention of the author to convince the reader that these circuits are exactly the same, but instead, that they are similar. Hence, the purpose of this work is to determine whether or not this similar circuit can be used for design. It is of some interest to consider the series-stack when there are more than two devices placed in series, this situation is analyzed in Appendix C, where it is found that N devices with $L = L_{min}$ in series yields an identical behaviour of a single device with $L = NL_{min}$. [1]

3.3 Op Amp Design Using Series Stack

Op amp design is application-specific. Due to the large number of variables, finding the optimum can be very complex, or in some situations, impossible. Hence, to find the

[1] For small-signal analysis of the series-stack, see Appendix D.

optimum to any problem, there must be well defined constraints. However, since this work is exploratory in nature, there is a lack of constraints and this makes the design strategy quite difficult. In the end, good approximations should be made to reduce the amount of free variables so that a local optimum can be found. It is important to note that the design strategy outlined below has been developed for this specific study, and is not intended to be an overarching method for all designs.

Optimization problems begin with outlining the variables as well as the constraints. The table 3.1 lists some major parameters that are defined based on the specifications, and other parameters that must be optimized.

Table 3.1: The operational amplifier design specifications

Parameter	Known	Unknown	Notes
C_L	10 pF		Determined to be around 10 pF when estimating test capacitances
C_C		X	This is usually set by noise, but there is no noise constraint here
Gain		X	No necessary target, but should be fairly high for testing purposes
Speed	100 MHz		Usually given by settling requirements, but here is chosen to be reasonable for testing
Feedback (β)	1		No target, so $\beta = 1$ was used as worst case estimate
Power		X	Should be minimized
Stability	$PM = 70^o$		Based on no overshoot, but also allowing for some process variation

By looking at table 2.1, it is clear that many details have been dropped to make the discussion manageable. However, many of the specifications outlined here are fundamentally important for integrated circuit design, so it makes sense to develop a strategy based on this.

As shown in table 3.1, many of the specifications are unknown. The problem is outlined by a short exchange between Alice and the Cheshire cat.

Alice: Would you tell me, please, which way I ought to go from here?

Cat: That depends a good deal on where you want to get to,

Alice: I don't much care where...

Cat: Then it doesn't matter which way you go!

To much freedom makes life more difficult. This highlights the challenge here, since most designs contain adequate specifications that force an amplifier into a certain direction; however, in this case it is difficult to determine where trade-offs should occur. As a result, extra analysis is needed to determine constraints such as the compensation capacitance.

3.4 Proposed Design Strategy

After the above discussion, a design strategy can be summarized. Note that this method is not meant to be general, but is specific to this application where the specifications are vague.

1. Determine needed gain, bandwidth, load capacitance and SNR for the application
2. Generate current density plots and choose bias point based on required gain and speed
3. Choose the amplifier topology and transistor length
4. Determine the required C_C to meet noise specification, if there is no noise requirement, choose $C_C = 0.2 * C_L$
5. Determine the required first stage transconductance based on Step 3 and the bandwidth/settling time
6. Adjust the size of the second stage to meet the stability requirement, while obeying $0.25 < j/k < 1$. [2] Note that C_C may be adjusted slightly to help reach the phase margin
7. Scale the entire amplifier to attain the desired bandwidth/settling time

[2] Large size mismatches between the two stages can cause signal propagation through C_{gd} of the second stage PMOS device, which will affect biasing.

Many of the design steps may seem confusing or not obvious. To alleviate this, a design example will be outlined showing the necessity of the steps, and a portion of the analysis used to determine these steps. It should be noted that this design strategy is of similar theme to that shown in section 2.3.2; however, there are a few changes due to the needs of this design.

The next chapter will now use the outlined design strategy, while providing insight into the analysis behind each step.

Chapter 4

Implementation

Here it is shown how to actually make an op amp, using the strategy outlined in the previous chapter. Each step will be discussed in order, and the discussion of the order will be touched on at the end. The technology used for design was the TSMC 65nm kit. This is not a FinFET technology, but should display similar square-law model breakdown, and give insight into the series-stack behaviour. It was also readily available with a relatively short turnaround between tape-out and delivery.

4.1 Design Flow

The order of steps in the strategy reflect typical design situations. In this work, some steps are actually reversed due to the lack of information. Specifically, the topology is determined initially, since it is the flagship modern analog circuit and well-understood. Hence, section 4.1.3 was the initial step for this work, but typical chronology is shown in this discussion.

4.1.1 Determine needed gain, bandwidth, load capacitance, feedback factor and SNR for the application

This step is strongly influenced by the given application. For this application, the gain does not have a specification but should be fairly high for testing purposes. The bandwidth is selected to be 100 MHz, since it is a fairly easy target to hit and reasonable to test. Furthermore, there is no noise specification, but the load capacitance can be estimated to

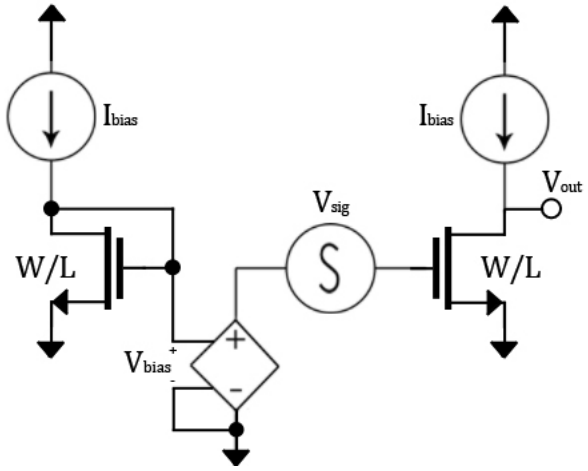

Figure 4.1: Setup to generate current density plots. V_{bias} is used to bias the gate of the amplifying transistor through an ideal voltage-controlled voltage source.

be roughly 10 pF (due to testing probes). Lastly, the future application is not known, so using a feedback factor of unity gives a worst case estimate.

4.1.2 Generate current density plots and choose a bias point based on required gain and speed

Generating current density plots is crucial in highlighting many fundamental limits of the devices, as discussed in previous chapters. Traditionally, parameters like the transconductance and output resistance have been plotted to highlight gain and speed. However, modern processes do not allow immediate access to such parameters, so a much more black box approach is taken here, where the gain and speed are plotted directly from simulation. The simulation set up is shown in figure 4.1. Note that the voltage-controlled voltage source has a perfect gain of unity, and it meant to buffer between the bias circuitry and the actual device. The widths of the devices are set to 1 μm for ease of calculation, and the current is then swept. The output node, V_{out} is then probed for gain, as well as DC voltage and bandwidth.

The current density plots of PMOS and NMOS devices can be found in figures 4.2 and 4.3, respectively. Note that the x-axis is in the units of $A/\mu m$, and that the shape is very different from the ideal Gm/Id plot shown in figure 2.6 – this is because these plots capture all the non-ideal effects of real devices. The plots were generated using the TSMC

65 nm kit, where the threshold voltage is estimated to be around 400 mV. It is important to realize that threshold voltage does not have an important role, but only serves as a guide to how the device is behaving (the fuzzy boundary between subthreshold, triode and saturation).

This project is focusing on devices that have short channel lengths, consequently, speed is not really an issue (100 MHz is attainable). However, DC gain becomes a concern, so the decision was made to pick the current densities in areas that have high gain (especially for the NMOS device). It is acceptable for the input PMOS to be in subthreshold, but the PMOS bias devices should be in saturation to properly mirror the current. This understanding leads to the following current density result:

$$\frac{i}{w_n} = 10\ \mu A/\mu m \tag{4.1}$$

$$\frac{i}{w_p} = 5\ \mu A/\mu m \tag{4.2}$$

$$\frac{i}{w_B} = 3.33\ \mu A/\mu m \tag{4.3}$$

The actual size of the devices is determined by layout concerns, knowing that $W < 20 * L$. The current densities selected have now already set a limit on the DC gain and the maximum speed. Based on the current density bias points, as well as figure 4.2 and 4.3, a gain estimation can be calculated. Noting that the gain simulated is for a single device, it is important to include the reduction of gain due to the load resistance. For the two-stage amplifier,

$$A_0 = 15.5dB - 3dB + 29.5dB - 6dB = 36dB \tag{4.4}$$

The subtractions are taking into account the anticipated load resistances of the bias circuitry, and the first stage only goes down by $3dB$ since the NMOS devices are twice the length of the input PMOS devices (due to noise concerns). Lastly, it appears that the bandwidth should be attainable based on the bias points selected.

4.1.3 Choose the amplifier topology

When choosing an amplifier topology, many factors need to be considered. Elements such as speed, gain, and power are all strongly affected by the circuit chosen. A very popular

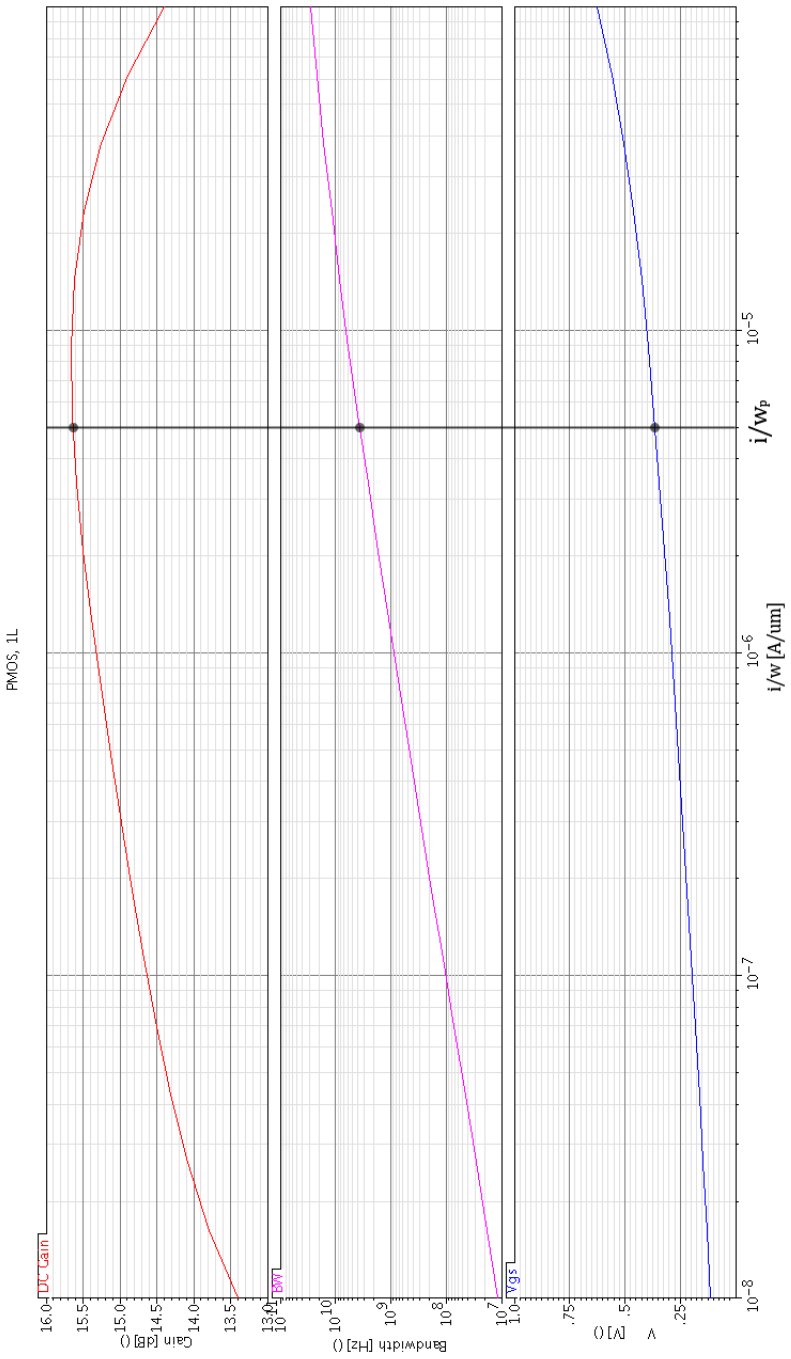

Figure 4.2: The minimum length PMOS current density plot.

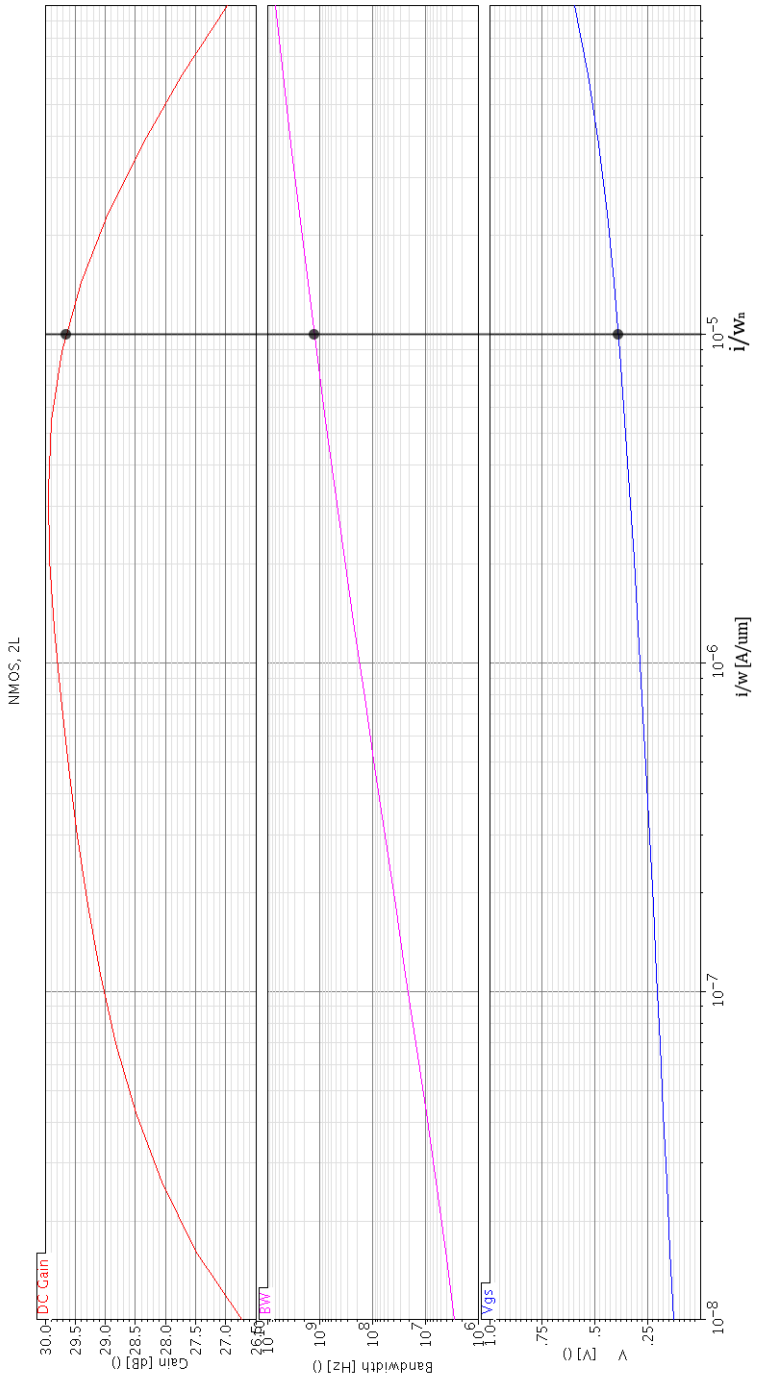

Figure 4.3: The NMOS current density plot.

choice for amplifiers is the two-stage op amp, since it delivers high speed, suitable gain and simplicity. If the gain specification of the design is high, a cascode topology may be chosen. However, the concern with cascode topologies are that in the sub-65nm technologies the supply voltage is around 1 V, and having a cascode eats up a lot of this voltage yielding a small signal swing. Furthermore, in practice differential circuits are preferred, but this adds complexity due to the required common-mode feedback circuit.

As mentioned in previous chapters, the current density plots allow the designer to determine the needed length, and topology in order to meet the gain specification. In this design, we are choosing to use the classic two-stage amplifier with PMOS input devices and a NMOS common-source second stage.

For noise concerns, the length of the NMOS transistors is made to be twice as long as the PMOS input transistors, specifically for the differential pair. The motivation for this is outlined in Gray and Meyer (fifth edition) p.781, where the flicker and thermal noise can be made dependant solely on the input device if the length of the NMOS is significantly larger than the PMOS (here significant is deemed a factor of 2). To review more about the noise and how it relates to the amplifier, see B.

At this point, the amplifier is mostly designed, and many of the performance specifications have been set. This is especially true for the DC gain, while the concerns left to deal with are settling (bandwidth) and stability (phase margin).

4.1.4 Determine the required Cc to meet noise specification, if there is no noise requirement, choose Cc=0.2*Cl

Choosing the compensation capacitor value is typically governed by the required noise level; however, if there is no noise specification then choosing C_C becomes difficult. To do this, it is best to focus on the location of the second pole and try to choose a optimum value for C_C.

Consider equation 2.15, if it is assumed that the transconductance is a constant, one can plot the pole frequency as a function of C_C. Note everything in the figure 4.4 is defined as a fraction of the load capacitance. The plot outlines at which point increasing C_C has diminishing returns on increasing the overall pole frequency. To determine the beginning of diminished returns, the point when the slope is unity has been highlighted. To cover the typical scope of parasitic capacitances, C_C can be chosen to be $0.2 * C_L$. Notice that this analysis is a slight approximation, since the actual value of g_{m7} will affect the slope of the equation; however, it is deemed a suitable approximation.

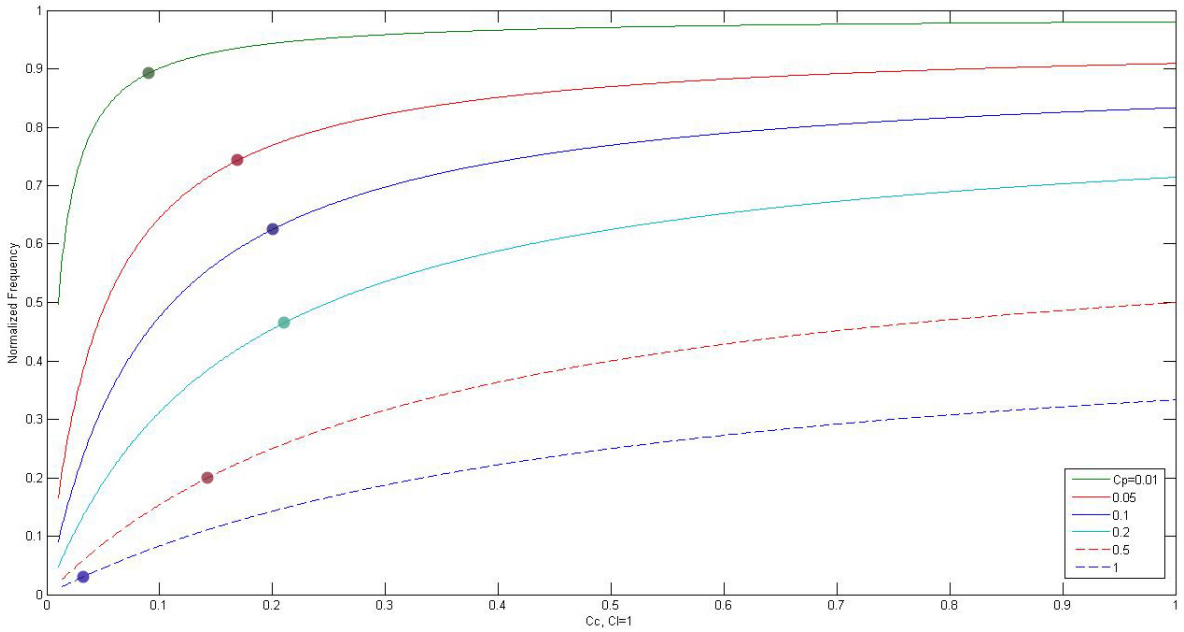

Figure 4.4: Normalized second pole frequency location.

An alternative approach to the second pole location analysis can be found in Appendix E, but note that the same conclusion is reached. Many designers may believe that choosing a compensation capacitor that is lower than the load capacitance could lead to an unacceptable phase margin. For most design methods this is true, but here, an additional adjustment is made: changing the relative size of the two stages. Sizing the two stages separately allows for increased stability, which will be shown later in this method.

4.1.5 Determine the required first stage transconductance based on Section 4.1.3 and the bandwidth/settling time

The required bandwidth for this work is 100 MHz and was selected for attainability and testability. To determine the first stage transconductance, the following calculations are used:

$$\omega_t = \frac{g_{m1}}{C_C} \tag{4.5}$$

$$g_{m1} = \omega_t * C_C \tag{4.6}$$

$$g_{m1} = 2\pi * 100\ MHz * 2\ pF = 1.26\ mA/V \tag{4.7}$$

At this point, the transconductance of the PMOS input device can also be calculated:

$$g_{mp} = \sqrt{2k'_p \left(\frac{W}{L}\right) I_{D1}} \tag{4.8}$$

$$g_{mp} = \sqrt{2k'_p \left(\frac{1}{L}\right) \left(\frac{I_\mu}{W_\mu}\right)} * jW_\mu \tag{4.9}$$

Where j is denoted as the first stage scaling factor, meaning it lists how many devices should be in parallel. All the variables are known, except for the technology factor k_p, but this can be estimated by doing some quick simulations (I_d vs. V_{gs} sweeps) shown in figure 4.5.

$$g_{mp} = \sqrt{2 * \frac{80\ \mu A}{V^2} * \left(\frac{1\ \mu m}{60\ nm}\right) * 5\ \mu A * j} \tag{4.10}$$

$$j = \frac{g_{m1}}{0.1\ mA/V} = 12.6 \approx 20 \tag{4.11}$$

This now gives an estimate to how large the first stage of the amplifier should be in order to attain the required bandwidth.

4.1.6 Adjust the size of the second stage to meet the stability requirement

As alluded to when choosing the size for C_C, the stability can also be controlled by changing the relative size of the two stages. This can be shown by considering the phase margin specification, and assuming that the system only contains two poles defined by equations 2.13 and 2.15. [1]

[1] The circuit contains additional poles, it is assumed that they are far beyond the unity-gain frequency.

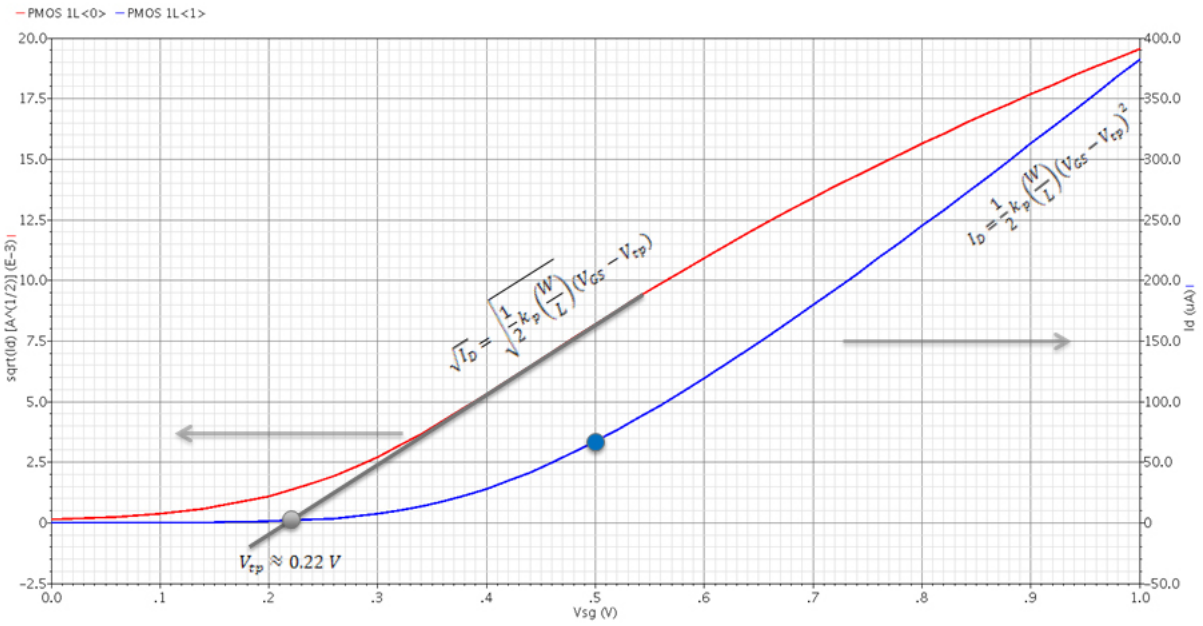

Figure 4.5: Drain current versus gate-to-source voltage for the input PMOS device, in order to estimate threshold voltage as well as $k'_p C_{ox}$.

$$PM = 70^o = 180^o - 90^o - \tan^{-1}\left(\frac{\omega_t}{\omega_{p2}}\right) \tag{4.12}$$

$$\frac{\omega_t}{\omega_{p2}} = \frac{1}{3} = \frac{k_1}{k_2} * \left(\frac{C_C C_P + C_C C_L + C_L C_P}{C_C C_C}\right) * \frac{j}{k} \tag{4.13}$$

Where k is a similar scaling factor as j but for the second stage of the op amp. Therefore, when both stages are scaled while keeping the ratio between them constant, there should be minimal change in the phase margin; however, one can adjust the phase margin through the relative ratio of j and k.

The second stage is now swept in order to determine the necessary size to obtain a phase margin of 70^o. Consider the plot shown in figure 4.6. To achieve a suitable phase margin, the second stage scaling k would have to be above 100. However, this violates the j to k ratio range [2]; thus, to hit the phase margin specification the compensation capacitor must be increased from 2 pF to 3 pF. The phase margin plot is then regenerated and it is found that, to obtain the phase margin specification, a $k = 80$ must be used.

This step is fairly straight forward; the j and k factors are increased while preserving their ratio until the bandwidth is achieved. In this case, that is when $j = 40$ and $k = 160$. Note that this step must sometimes be repeated since after scaling, the phase margin may fall below the threshold, causing one to increase C_C or decrease the j/k ratio. Obviously, one cannot scale the stage infinitely to reach infinite speeds; the limit comes when the second pole approaches the higher order poles due to the differential pair mirror.

4.1.7 Scale the entire amplifier to attain the desired bandwidth/settling time

At this point the op amp is fully designed, as shown in table 4.1, and one can now commence schematic testing to ensure that design specifications are met and performance is reasonable. Note that the design discussed in this section will be used for all amplifiers. This means that all current densities, scaling factors and compensation capacitors will be kept the same; the only changing parameter in this work is the length (and length implementation) of the devices. [3]

[2] The stage ratio, j/k, must be larger than 0.25 to avoid coupling of the output to the bias transistors. Specifically, there is concern that the output will flow through C_{gd} of Q_6 in figure 2.4.

[3] It turns out that the compensation capacitor did need to change for 2L designs, in order to meet the specifications.

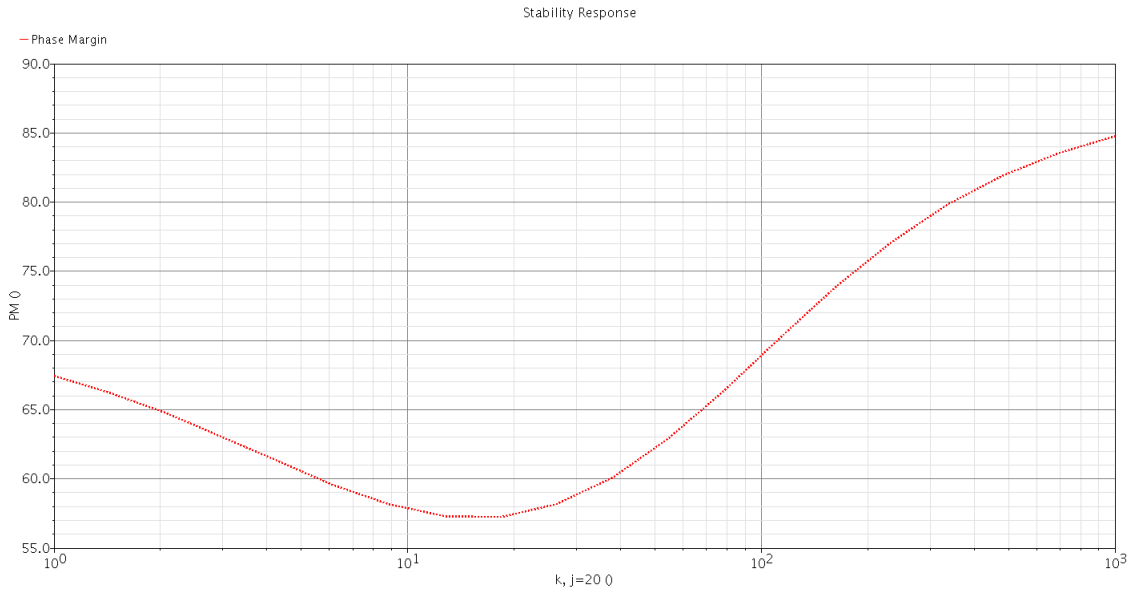

Figure 4.6: The j to k ratio (scaling of first and second stage) being scaled with a fixed j.

A complete schematic with all (W/L) values is shown in Appendix F, specifically for the 1L design.

4.2 Op Amp Schematic

To test the series-stack, four amplifiers were designed. Since the non-idealities of short devices pose challenges to designers, the two test lengths were L_{min} and $2L_{min}$. Both topologies were implemented with traditional bulk transistors and the series-stack resulting in four amplifiers. Each design will be tested at the schematic level, highlighting important differences and effects, but the discussion will be left for the end of the chapter.

Implementing the parameters outlined in table 4.1 is trivial; but now the discussion moves towards testing. The methodologies for testing the op amp at the schematic level are found in Appendix G. Here, the final results will be summarized, while giving clarification where the author deems it is necessary. The final results of all amplifiers are shown in table 4.2. Note that the output swing is measured at the voltage when the gain drops by half of its maximum value.

After inspecting table 4.2, the reader should first realize that the phase margin is not

Table 4.1: Op amp design parameters

Parameter	Value
C_L	10 pF
C_C	3 pF
I/W_n	10 $\mu A/1$ μm
I/W_p	5 $\mu A/1$ μm
I/W_B	3.33 $\mu A/1$ μm
j	40
k	160
V_{DD}	1V

set at the 70^o mark as discussed previously. This is a retroactive change, which will become apparent in the layout results, and will be left to the discussion section. Overall, the results are reasonable, with significant changes in gain and speed as the length is doubled, while the two implementations for a constant length remain comparable.

A more in-depth analysis will be left to the discussion section. However, the reader is encouraged to be familiar with the table for future reference, and also keeping it in mind while viewing the layout results.

4.3 Op Amp Layout

Transforming the circuit abstraction into a physical device is the goal of layout. Integrated circuit layout can be very complicated, and difficult to explain since there are numerous decisions and variables influencing the process. To highlight the key aspects of layout in this work, general rules of thumb can be presented. These rules were used throughout the design to ensure consistency and good performance.

- Transistor widths were chosen based on the following constraint: $W < 20 * L$. This ensures low polysilicon-gate resistance
- All transistor gates were parallel
- Differential pair transistors were divided into groups, and laid out in a common-centroid fashion

Table 4.2: Op amp schematic results, TSMC 65nm, 1V Supply

	L_{min}	S-L_{min}	$2L_{min}$	S-$2L_{min}$
DC gain	37.2 dB	35.2 dB	49.25 dB	49 dB
Unity-gain frequency	47.26 MHz	43.02 MHz	26.25 MHz	21.54 MHz
Phase margin	83^o	88^o	84.4^o	94^o
Settling time, 100mV Step	20 ns	20 ns	35 ns	36 ns
Systematic offset	2.5 mV	1.84 mV	0.43 mV	0.9 mV
Sys. off. variance	0.8 mV	0.65 mV	0.5 mV	0.84 mV
Output swing	760 mV	760 mV	695 mV	662 mV
CMRR	39.32 dB	36.2 dB	47.8 dB	47 dB
Power dissipation	1.5 mW	1.35 mW	1.54 mW	1.21 mW
Equiv. input noise, 100 kHz	20 $nV/\sqrt{Hz}$	18.7 $nV/\sqrt{Hz}$	13.2 $nV/\sqrt{Hz}$	13.56 $nV/\sqrt{Hz}$

- Wire widths were determined using expected capacitances and calculated metal layer resistance
- All capacitors were implemented as metal-insulation-metal (MIM) capacitors, for simplicity and potential area conservation in future designs
- Biasing was accomplished using a current source, which is planned to be implemented at the testing level to ensure proper bias current
- Parallel transistors were implemented using multiple gate fingers
- Transistor source resistance was designed to be constant among transistors, to ensure constant V_{gs} for the bias point

The layout design for the bulk transistor configuration, based on the general rules is shown in figure 4.7. Similarly, the layout for the series design is shown in figure 4.8. Although not obvious from the figure, the series-stack offers a "cleaner" layout. Because of the organizational advantage, the design can be more compact compared to the bulk design. This is due to the fact that transistors in each stage have the same scaling constant, so all devices line-up perfectly. A view of the completed chip, with all four amplifiers is found in figure 4.9.

Table 4.3 summarizes the post-layout extraction results. The results will be analysed further in the next section. At this point, it is worthwhile to note the significant change in PM, which necessitates the revised PM in the schematic design.

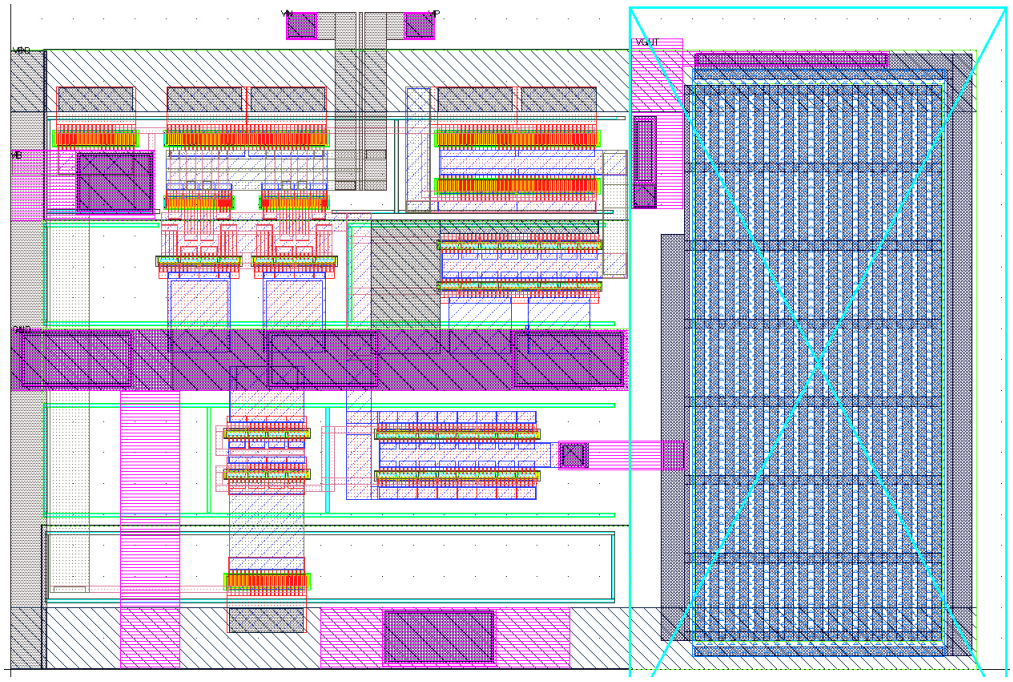

Figure 4.7: The bulk L amplifier layout.

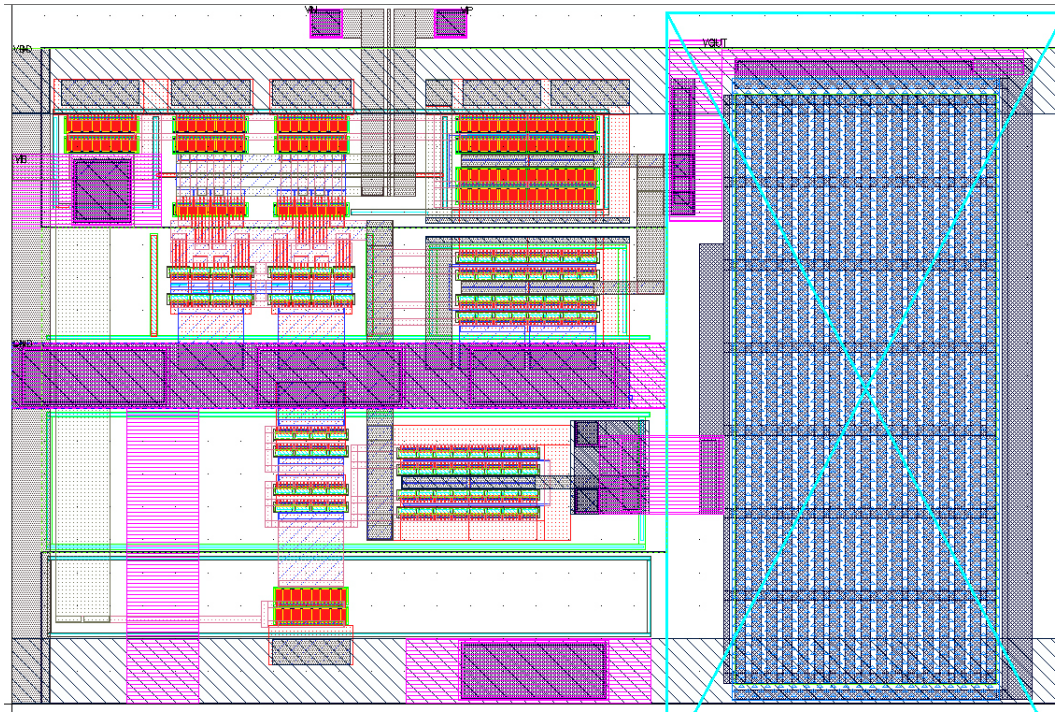

Figure 4.8: The series-stack L amplifier layout.

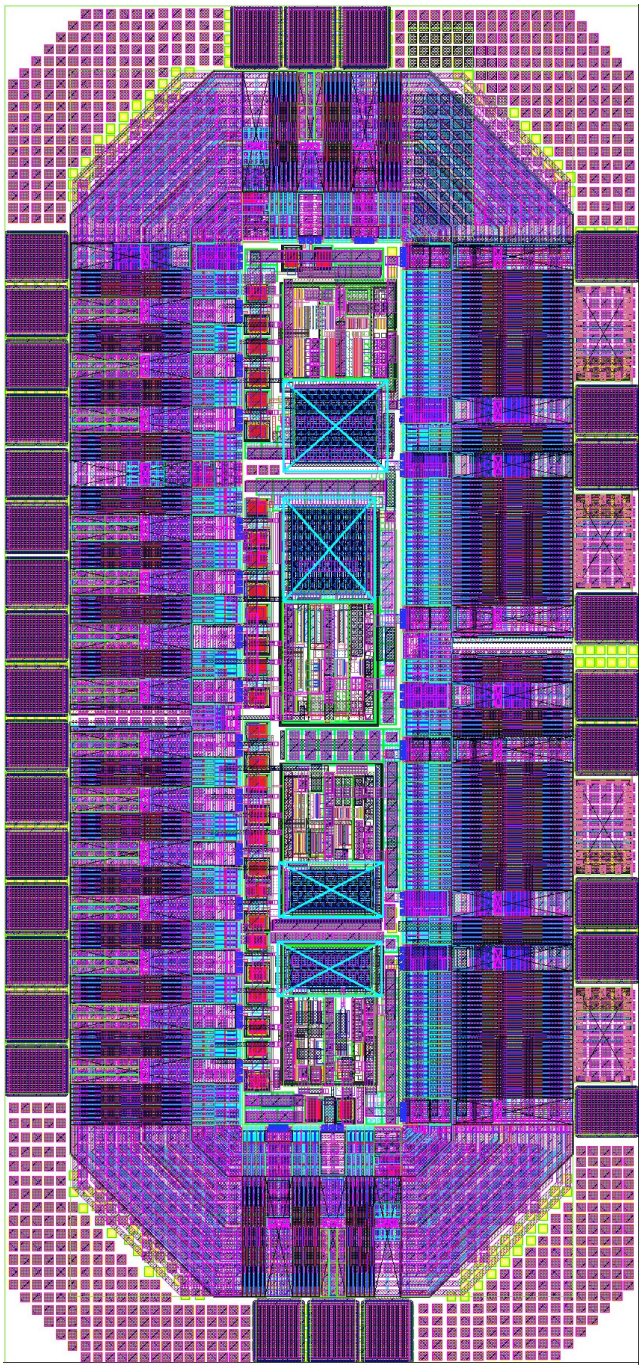

Figure 4.9: The completed chip.

Table 4.3: Op amp layout results, TSMC 65nm, 1V Supply

	L_{min}	S-L_{min}	$2L_{min}$	S-$2L_{min}$
DC Gain	37.45 dB	37.14 dB	48.5 dB	52.3 dB
Unity-gain frequency	45.2 MHz	37 MHz	32 MHz	37.8 MHz
Phase Margin	68^o	67.2^o	69.4^o	63.8^o
Settling time, 100 mV Step	30 ns	30 ns	36 ns	40 ns
Systematic offset	0.36 mV	0.47 mV	0.125 mV	1 mV
Sys. off. variance	0.97 mV	1.3 mV	0.3 mV	0.72 mV
Output swing	806 mV	740 mV	680 mV	640 mV
CMRR	42.7 dB	42.7 dB	57.3 dB	54.75 dB
Power dissipation	6.25 mW	5.2 mW	4.63 mW	5.04 mW
Equiv. input noise, 100 kHz	4.26 $nV/\sqrt{Hz}$	4.1 $nV/\sqrt{Hz}$	3.6 $nV/\sqrt{Hz}$	3.1 $nV/\sqrt{Hz}$

4.4 Results Discussion

To ensure a structured analysis, the discussion will be divided into comparison pairs. It is important to realize that there are many metrics that could spark interesting conversations, but only those that have a significant impact on this thesis will be discussed. Any other realization will be left for future work.

4.4.1 Schematic: Lmin vs. S-Lmin

Taking a look at the schematic results, and comparing L_{min} and $S-L_{min}$, key difference in gain, speed, phase margin and systematic offset are observed. Gain and systematic offset are both DC parameters, indicating that the DC behaviour of the amplifiers is different. The key here is to ensure that the DC difference is due to the fundamental modelling. To test this, a simple schematic was used, implementing the circuit diagram shown in figure 3.1.

The resulting DC operating points show that there is a model difference between the two implementations, shown in figure 4.8. Note that the devices proved to be the same using square-law models, but the simulation software is more sophisticated. Therefore, this explains the discrepancies in DC gain and systematic offset.

Turning to the AC performance, the differences in speed and phase margin to be examined. By simply considering the relative values of the unity-gain frequency, the difference

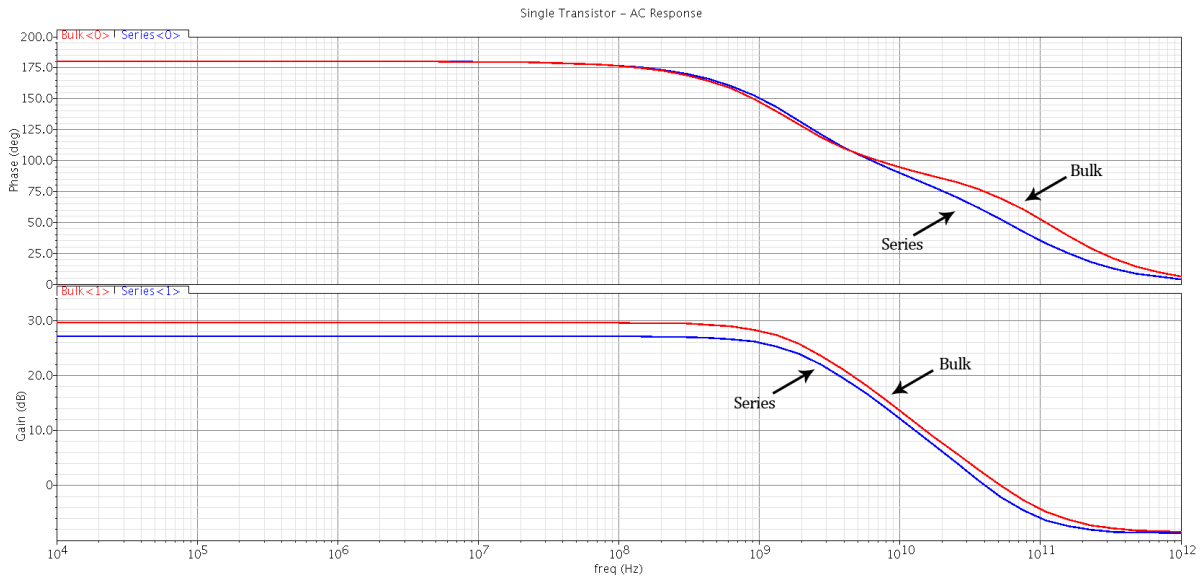

Figure 4.10: The bulk vs. series amplifier schematic results.

in phase margin is expected. [4] This hints at the fundamental AC difference – the unity-gain frequency. However, this may be easily investigated by the same circuit used for the DC discussion. It is expected that there will be a difference in the AC performance since the series stack introduces a capacitance in the middle of the channel. In simulation, it is apparent that the series implementation is fundamentally slower, by about a factor of three.

Using the conclusions found in the analysis above, almost all the discrepancies in the schematic results can be explained.

4.4.2 Schematic vs. Layout: Lmin and S-Lmin

When comparing schematic vs. layout results, it is useful to examine all designs, it will become apparent that many trends occur in all designs. To begin, the first design, L_{min}, will be examined. Immediately, differences arise in speed, phase margin, systematic offset, and power consumption. Many of these differences can be explained by acknowledging the introduction of parasitic resistances and capacitances. This holds true for the $S - L_{min}$

[4] The series-stack is slower, meaning additional poles approach the unity-gain frequency, resulting in a phase shift increase which leads to a lower phase margin.

design as well. However, when looking at the $2L_{min}$ and $S - 2L_{min}$ designs, the results become unexpected and will be discussed in the next section.

4.4.3 Schematic vs. Layout: 2Lmin and S-2Lmin

Going from schematic to layout, it is expected that the circuit becomes slower; however, with the $2L_{min}$ and $S - 2L_{min}$ implementations, this is not the case. To investigate this, a simpler circuit will be used to pin point the issue (it turns out that the operational amplifier is quite complex, and finding the root of this problem is challenging). To try and reproduce this phenomena, a simple inverter was designed with the specifications shown in table 4.4. The values were chosen to best reflect the devices used in the amplifier design. Since the comparison is being done between schematic and layout, the laid out inverter is shown in figure 4.11. Again, the devices were laid-out in a manner consistent with the transistors in the amplifiers.

Table 4.4: Test inverter specifications

Parameter	Value
W_p	1 μm
L_p	130 nm
W_n	500 nm
L_n	130 nm
Fingers	40

To begin the comparison, consider the DC sweep results shown in figure 4.12. Due to the inherent gain of the configuration, the DC bias point can significantly change; hence, for the rest of the analysis, the common-mode voltage at the input is set so that the output voltage remains half-way between the rails. Using this, an AC sweep can be performed, where the results are plotted in figure 4.13. The AC sweep provides great insight into the trends observed in the amplifier results, mainly that the layout phase performance is significantly worse than the schematic and there are two contributing effects causing this. Firstly, considering the phase response, it is clear that although the schematic phase seems to settle after the first pole, the extracted results do not exhibit this behaviour and continues to decrease. Secondly, the first pole of the schematic result occurs before the extracted result, meaning that the extracted circuit is faster than the schematic. This second point is very concerning, and calls into question the simulation software as well as the models being used since the extracted results should almost always be slower than

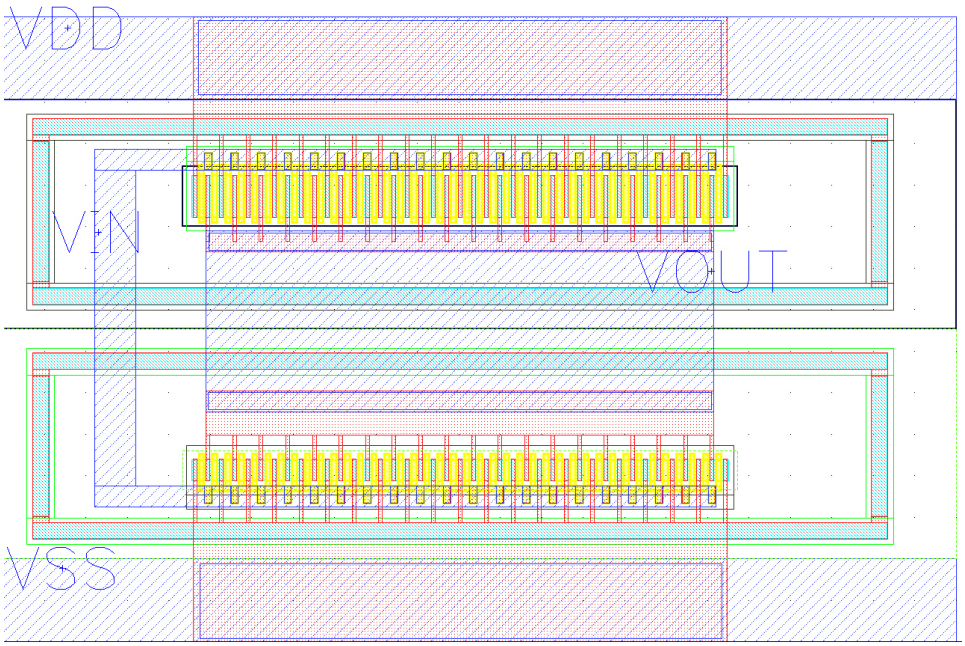

Figure 4.11: The test inverter layout.

the ideal schematic. However, this inverter simulation shows that the amplifier results are consistent with the device models, and there isn't necessarily an issue with the circuit design but something more fundamental.

A hypothesis that could explain the large discrepancy in speed, could be that the simulation software is overestimating the parasitic capacitance and resistance related to the devices. It has been mentioned that the transistors were laid out in fingers; hence, certain capacitances and resistances were decreased - perhaps the software is not taking this into account. To test this, an additional inverter was designed, using two big transistors. The DC sweep is shown in figure 4.14, where the previous result is overlaid for comparison. Notice that there appears to be a significant simulation difference when using a fingers implementation versus the bulk transistor. This signifies that the model does not take into account the resistive savings of laying out transistors in a finger topology. Additionally, an AC sweep was also simulated and is shown in figure 4.15. In a similar fashion to the DC analysis, it would appear that the models at the schematic level are capable of giving good predictions when using conventional bulk transistors, where any differences in speed are expected (the layout version is slightly slower due to additional parasitics). Referring back to figure 4.13, this realization can be used to pin point the problem, where

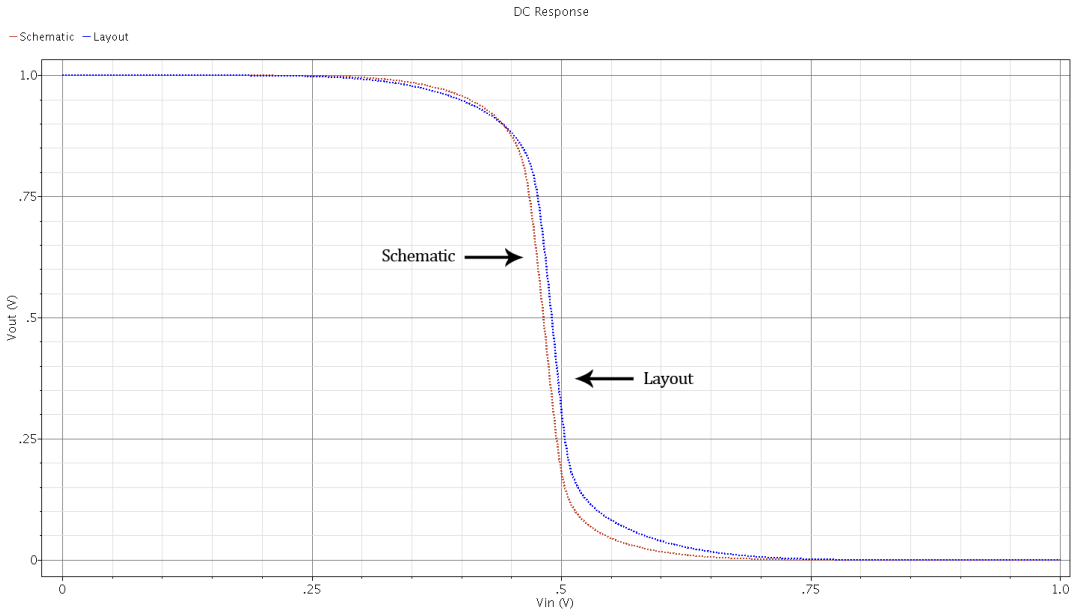

Figure 4.12: DC sweep of the test inverter.

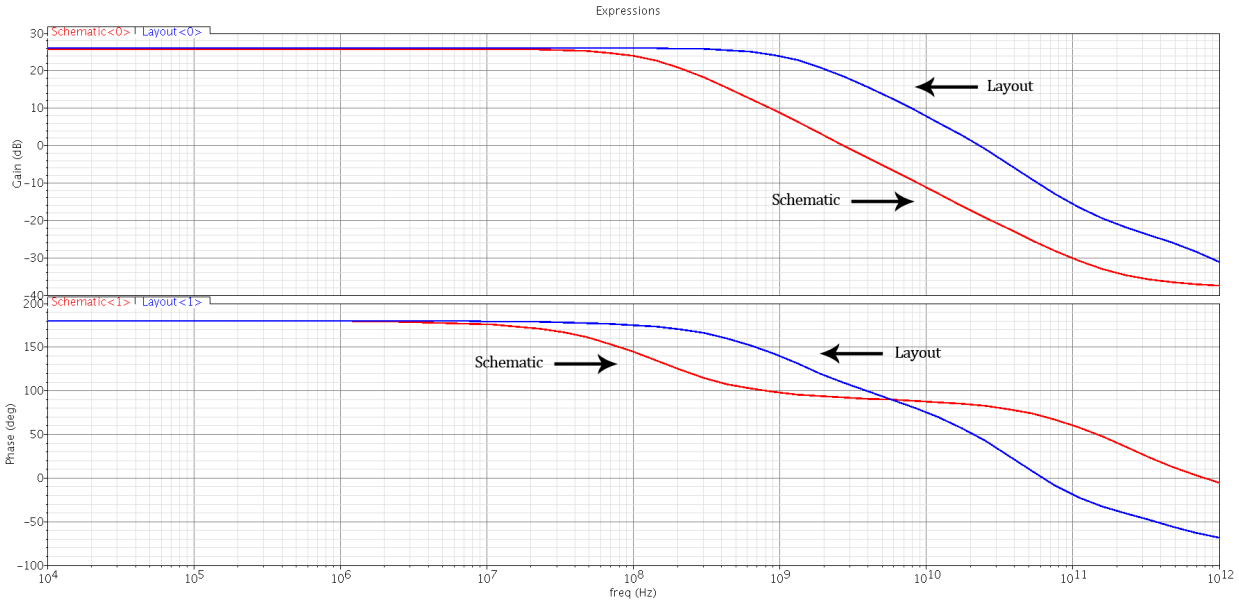

Figure 4.13: AC sweep of the test inverter.

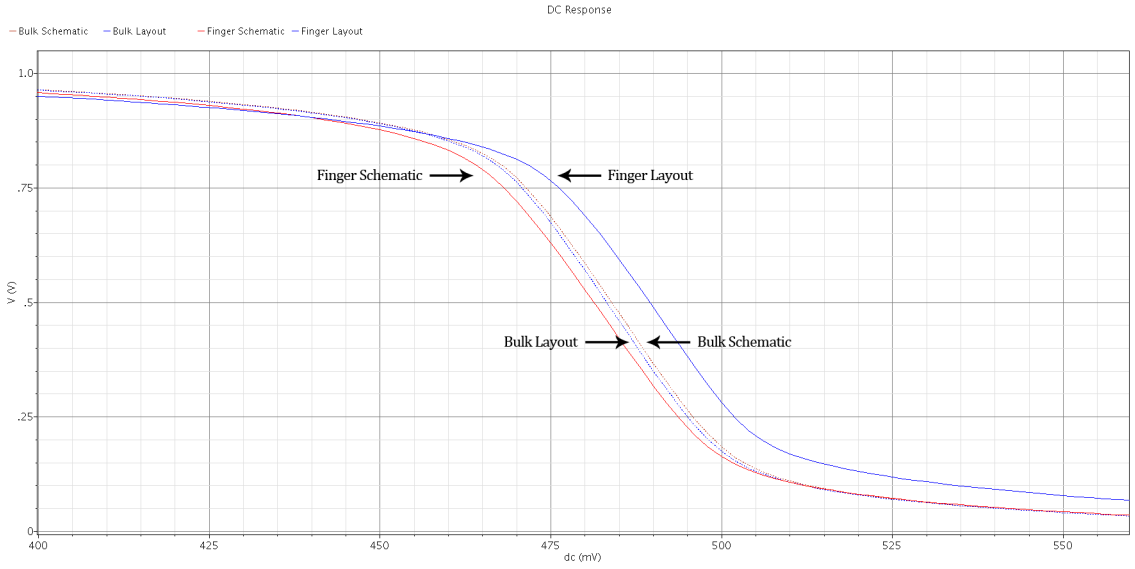

Figure 4.14: The inverter DC sweep with big transistors.

the models do not properly take into account the finger topology. At this point, the reader may begin to think that the accountability for the inconsistency belongs to the designer; however, as depicted in figure 4.16 the simulation software has a specific field for finger transistor implementations. Hence, it is assumed that this would be taken into account at the schematic level, but it isn't - yielding the problem observed here.

Solving the finger simulation issue is not straight forward, since it really entails redefining models, or the entire strategy. Hence, investigations on how to solve the problem encountered here will be left for future work. Possible solutions will be presented in the next section. Before moving on, the actual result of interest should be discussed. It is interesting that the main difference between designs is likely caused by problems in the models for transistor width control (finger vs. bulk) while the actual investigation was to consider length control (series vs. bulk). Although it has been mentioned that the differences between bulk and series length implementations can be explained through simulation model limitations, it does not answer the question of whether or not this is a viable design strategy.

Considering the results shown in tables 4.2 and 4.3, the general treads are equivalent between the bulk and series topology. Both topologies show similar behaviour when changing length for the gain, and unity-gain frequency. Additionally, the noise is also decreased as the length increases, which is expected. It it not enough to simply have similar be-

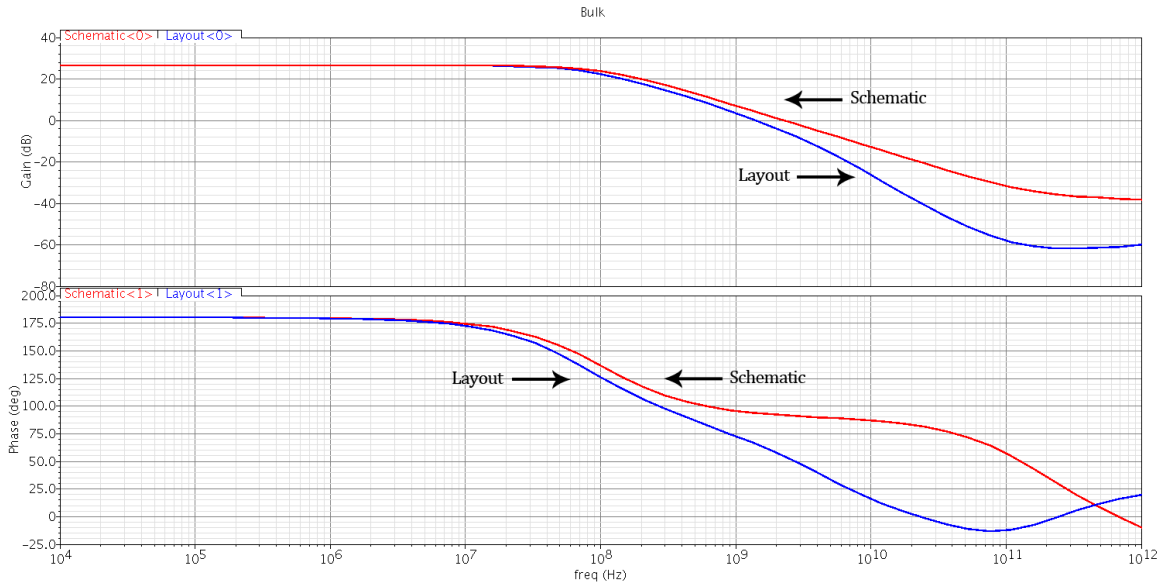

Figure 4.15: The inverter AC sweep with big transistors.

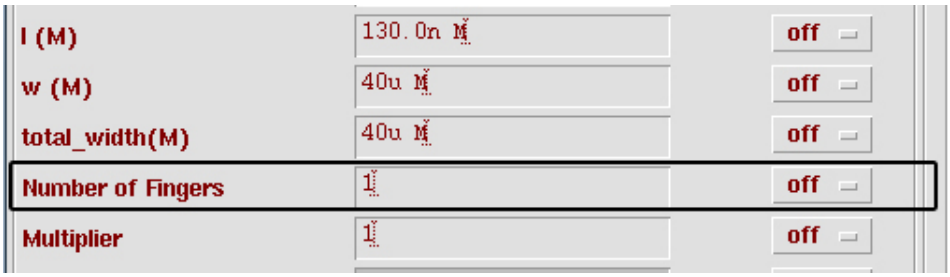

Figure 4.16: Simulation software input box.

haviour, the designs must also attain similar values, and this is observed in the results. From the schematic results, both gain and unity-gain frequencies are similar, with an error margin of about 20%. However, this error is somewhat expected due to the differences in parasitic capacitance between the designs, specifications like gain, CMRR and noise remain very similar (error $\approx$ 2%). Lastly, other DC driven properties like systematic offset and output swing are significantly different (error = 40%); yet, these types of specifications are typically not determined beforehand on paper, and are simply a result of simulation to be tabulated. Thus, the differences do not pose problems to using the series-stack design, they must simply be noted.

In conclusion, by neglecting the simulation discrepancies between schematic and layout, using a series-stack of transistors to behave as a longer bulk transistor does seem to be a viable approach for amplifier design. It should be noted that some of the specifications will change, but the trends as current, width and length are scaled remain the same. Hence, if a designer takes the approach of using current density plots to first characterize the devices, the series-stack method is favourable. The advantages of using the series-stack is to operate within a handful of well-modelled devices and having a cleaner layout design. The disadvantages are some changes in DC operating points, where specifications can be different than expected. However, the author believes that this is a suitable compromise. Based on these findings, the series-stack seems like an advantageous method for design; pending fabricated chip results.

4.4.4 Refined Design Strategy

The typical design strategy has led to poor results, due to errors in the simulation models. To outline a more accurate strategy, the problem should be properly defined. Consider the first pole location of the amplifier, as shown in equation 2.13. Now, the reasoning presented previously for the inverter performance exhibiting faster layout behaviour was due to overestimation of capacitance by the simulation software. However, in the case of the op amp, the first pole does not depend on any device capacitance, but solely on C_C. Thus, the reasoning for the model discrepancy is not the same as the inverter case, but there are other factors that can be affected by layout effects. A more detailed figure for the results in $S - 2L_{min}$ (since this design topology exacerbates the issue) is presented in figure 4.17. Terrible effects happen around 300 MHz, but this is somewhat expected due to all the high frequency poles and zeroes. The important factors are the dominant pole location, and the phase behaviour near the unity-gain frequency.

In the case of the schematic, the phase actually increases near the unity-gain frequency. This is an effect that occurs when the compensation capacitor has a large value, making

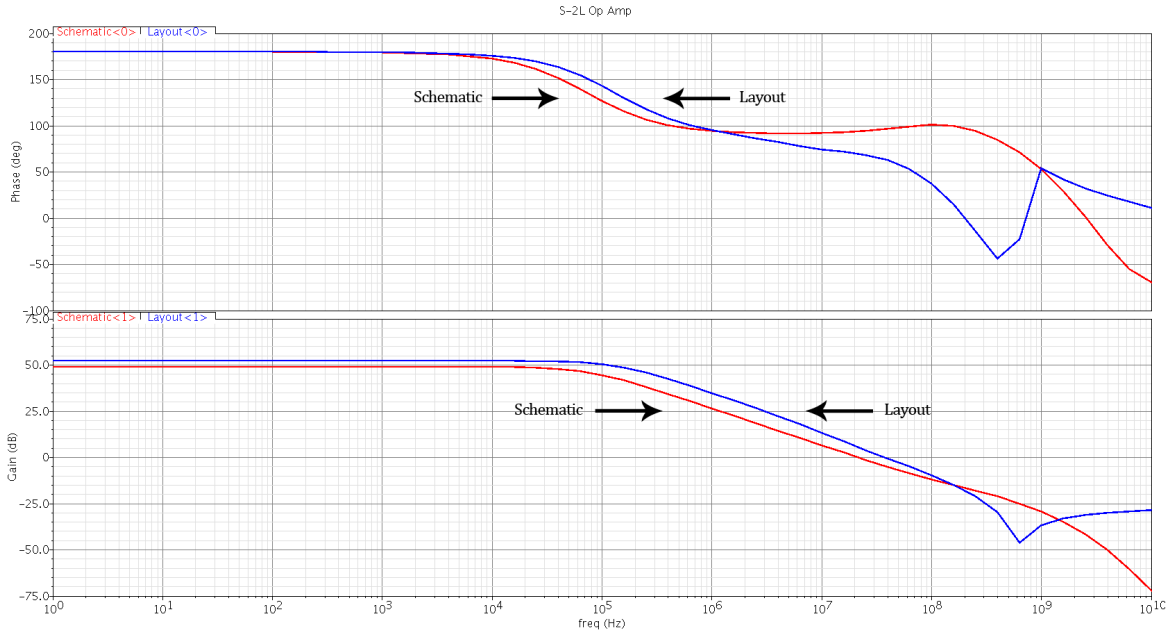

Figure 4.17: AC sweep result of S-2L amplifier.

the zero approach the frequency of interest. Note that typically this would not be an issue, since the compensation capacitor was solely increased to alleviate the discrepancy here. Turning the attention to the layout plot, one can see that the amplifier is not only faster, but exhibits a disparity in phase behaviour. The combination of these two factors leads to problematic phase margin differences. Although it is believed that the difference in speed is due to overestimation by the simulation software, the difference could also be due to DC biasing. The DC biasing problem is apparent in the power dissipation discrepancy between schematic and layout. With these factors in mind, it is clear that the simulation software at the schematic level cannot be relied upon when using a finger topology, since it calculates capacitance and resistance incorrectly.

An immediate possible solution for the schematic to layout discrepancy could be to simply base design decisions off of the layout results, since these are assumed to be closer to reality. Such an approach would yield reasonable results. However, having to lay out every single test device would be very time consuming, which is the opposite of what is trying to be accomplished with the strategies presented in this work.

Another possible solution would be to attach negative capacitors to the nodes of the circuit. The simulation software should not have large concerns regarding the negativeness

of the capacitors, and they could then be tuned to better match the layout results. This method would allow the designer to work in the schematic environment, which is very advantageous when doing width-scaling sweeps. Difficulties would arise when trying to properly tune the capacitors; however, once tuned for a specific size, this method could be robust.

4.5 Testing

Many of the results found in simulation do not hold much weight if the measured results do not agree. Hence, it is paramount that the chip be tested. The timeline of this work does not allow the test results to be included. Nonetheless, a thorough testing scheme will be included for completeness. Many of the amplifiers' specifications were chosen to ease testing, for example, the gain was chosen to be large but not too large that it would pose problems in testing (properly identifying the large gain values can be challenging). Although the frequencies of interest are not exceedingly large, they are large enough to introduce problems for breadboards - meaning that the testing of this chip will require a printed circuit board (PCB). This leads to the first section of testing, design a suitable PCB.

4.5.1 PCB Design

The envisioned design for the printed circuit board is not complicated, but simply requires a suitable interface connection between the packaged chip and the co-axial cables. Although it may be interesting to include other complicated features like digital switching between amplifiers, this is not required and may simply lead to increased failure modes. Instead, a simple PCB is proposed to connect each amplifier to their own co-axial cable. Note that each amplifier has five ports: V_{i-}, V_{i+}, V_{out}, VDD and VSS. On the chip, the distinctness of each amplifier is respected, but at the PCB level, this becomes impractical since that would lead to 20+ co-axial connections. Instead, global power connections will be used, while keeping the inputs and outputs distinct. Such an implementation lowers the amount of co-axial connections to 15 (three inputs/outputs per amplifier, and three global power nodes). The connection map is summarized in table 4.5 below.

At this point, testing simply becomes a question of connecting to the appropriate ports, and ensuring proper measurement technique, which will be discussed in the next section.

Table 4.5: This table shows the connection map between the packaged chip and the PCB design, specifically illustrated for the first amplifier.

Parameter	Chip Port	PCB Port
V_{i+}	Op Amp 1L	Op Amp 1L
V_{i-}	Op Amp 1L	Op Amp 1L
V_{out}	Op Amp 1L	Op Amp 1L
VDD	Op Amp 1L	Global
VSS	Op Amp 1L	Global
IO VDD	Chip	Global IO
IO VSS	Chip	Global IO

4.5.2 Power-on Cycle and Measurement

Before attempting to measure the amplifiers performance, a specific power-on routine must be followed as outlined by the chip manufacturer, TSMC. Note that the power-on sequence is dictated by IO cells within the chip, which control ESD circuits as well as power to the IO ring as a whole. Hence, it is important that these steps are followed exactly to ensure the electrical safety of the chip. The power-on cycle is outlined below:

1. Turn on the higher (I/O) voltage
2. Turn on the lower (core) voltage
3. Perform needed measurements and tests
4. Power down the lower (core) voltage
5. Power down the high (I/O) voltage

Following this cycle ensures that the power-on control cell is being properly utilized. Note that for this technology, the higher voltage (IO) is 2.5V, while the lower voltage (core) is 1V.

Gain, Speed and Phase Margin

As mentioned, the gain was designed to ensure testability. Usually, high gain amplifiers can be difficult to measure, but in this case a simple open-loop gain can be measured. The

input-signal should be sufficiently small to ensure the output does not swing significantly, while it should be large enough to avoid noise distortion. The input frequency can then be tuned while measuring the output to measure the unity-gain frequency as well as the phase. The frequency is not exceedingly high to where it becomes problematic to measure on an oscilloscope; hence, simple measurements should be sufficient.

Settling Time

In a similar fashion to the gain measurements, the settling time can be measured readily by using a setup similar to that used in simulation. The amplifier should be connected in a unity-gain configuration. Then a small-signal step function can be applied to the input while measuring the output.

DC Parameters

Measuring the DC parameters is slightly different than done in simulation since their is no voltage-controlled voltage sources. Yet, the actual measurement is straight forward. A differential signal with a common-mode can be applied to extract the voltage offset. Additionally, a small-signal source can be used while varying the DC input bias to determine the output swing.

Common-mode Gain

Measuring the common-mode gain is simply a matter of applying the small-signal to both inputs instead of just one, as in the gain measurement case. The CMRR is then taken to be the ratio of both gains.

Testing of the real circuit is of paramount important, since this reflects what is actually going on in the devices, instead of simulation software interpolating between nodes. Hence, once testing has been completed, it will be compared to the schematic and layout results, and final modifications to the design strategy will be performed. It could be the case that the chip exhibits similar behaviour to that of the schematic results, which would lead to a very similar design strategy. However, it is more likely that the actual chip will exhibit novel behaviour, leading to more insight into the design process and simulation models.

Chapter 5

Summary

It is found that there are fundamental simulation differences between the bulk implementation and the series-stack devices. However, this work does not consider this a serious concern, since the design strategy is based on simulation plots which will reflect the discrepancy. The key concern to highlight was whether or not the series-stack obeyed similar fundamental trends with current density and length. For the majority of tests, this seems to be the case. Not only this, but the actual relative values between the series-stack and traditional bulk devices were similar, with a typical error of 10%. With similarities in performance, the largest difference occurs at the layout level. The series-stack is more organized, since transistors in the same stage are scaled by the same factor leading to a very predictable size. An increase in organization translates into more compact designs. This advantage alone justifies the slight performance discrepancies and the price in speed that the series-stack requires (as expected, the series-stack tends to be slower).

A significant downfall of the design strategy is found when trying to predict speed and phase margin at the schematic level. This problem manifests due to shortcomings in the simulation software (Cadence was used here). Specifically, the software did not take into account the capacitive and resistive savings when laying out devices in multi-finger configurations. This leads to underestimations in speed, which after layout, causes significant decrease in phase margin. The decrease in PM can also be observed in the transient response. A solution to such software complications are not obvious; hence, new ideas are needed here. It should be mentioned that the series-stack does use more signal-swing to operate; this is observed in simulation. Additionally, the gain does tend to be lower with the series-stack, with an error of 5%. The additional node in the middle of the channel causes measurable differences in AC performance, the series-stack is 15%

slower than the bulk devices and has worse phase performance. Hence, to implement the series-stack, the designer must be aware of the costs in speed, signal-swing and stability.

Future work should include testing the actual chip to evaluate model accuracy. To help with this, a testing plan was outlined. The test results will give valuable insight into whether or not the series-stack implementation can be used for future design methodologies, specifically, whether or not the models properly reflect fabrication silicon. In order to rectify the discrepancy between schematic and layout, an inquiry into the model libraries should be undertaken. If this proves unsuccessful, other means outlined solutions should be investigated like the negative capacitor approach.

Overall, the series-stack solution to the fixed channel length problem is a viable option for future designs. Although exact estimations of circuit performance may not be straight forward, general design trends still hold. For example, the relationships between length and current density with gain and speed (f_t). The most apparent advantage of the series-stack is not easily shown in documentation, like this report. Its main benefit is at the layout stage, where organizing unit transistors of similar length becomes trivial and leads to clean and simple designs. Such an advantage is worth the slightly unorthodox nature of the series-stack, especially since the behaviour with respect to length remains similar.

APPENDICES

Appendix A

Cascoding

Here, a slightly more involved analysis will be done regarding the cascode topology. Considering figure 2.9, the small-signal model can be drawn and is shown in figure A.1. Using typical Norton analysis, the gain can be determined by finding the output resistance and transconductance. The transconductance can determined almost by inspection, being g_m. Determining the output resistance is slightly more involved, but by shorting the input to ground, the following result is found:

$$R_{out} = r_o + R_s + g_m r_o R_s \tag{A.1}$$

Where R_s is the resistance in the common-source transistor, usually r_o. In most cases, the $g_m r_o$ is much larger than both resistances, so the equation reduces to the following:

$$R_{out} \approx g_m r_o R_s \tag{A.2}$$

Putting the equations together...

$$A_o = g_m[g_m r_o r_o] \tag{A.3}$$

At this point, it should become clear why cascoding increases the gain by a factor of $g_m r_o$. This is a greater increase than would be accomplished by simply doubling the length or putting another device in series, as this work discusses. Note that the cascode topology effectively boosts the output resistance of the amplifier.

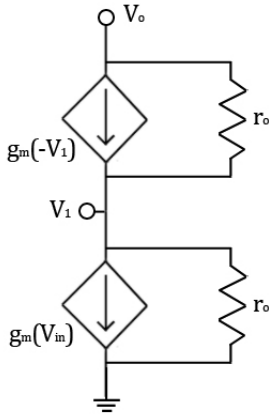

Figure A.1: The cascode small-signal model.

However, there are trade-offs to consider, these involve increased noise and extra biasing overhead. A thorough review of noise theory is shown in Appendix C, but for the purposes of discussion the results show that the noise of a cascode circuit is governed solely by the size of the common-source transistor. This means, in comparison with the series-stack, that although the gain of the cascode yields greater values, the noise performance is worse.

The overhead consideration is that the additional transistor must be biased. To set the DC voltage of the cascode transistor, a bias circuit must be designed which creates an additional path to ground. Additional biasing consumes additional power.

Appendix B

Noise Analysis

A noise discussion has not been the focus of this work, but in this section, it will be quickly reviewed to better understand design choices and why the series-stack is a favourable topology in the context of noise performance. The analysis will begin with individual transistors, then move to the two-stage CMOS amplifier topology used in this work.

B.1 Transistor Level

Transistors are active components which can be characterized by two types of noise sources: flicker (also known as 1/f) and thermal noise. Although interesting, the exact physical reasons to why this noise arises is not important for this work, but there sources are typically as shown in figure B.1. Using device physics, it is possible to derive equations that characterized these noise source as a function of the device parameters, where the flicker noise can be written as:

$$V_g^2(f) = \frac{K}{WLC_{ox}f} \tag{B.1}$$

Where K is a parameter dependant on device characteristics. The thermal noise can also be derived as (assuming active region operation):

$$I_d^2(f) = 4kT\gamma g_m \tag{B.2}$$

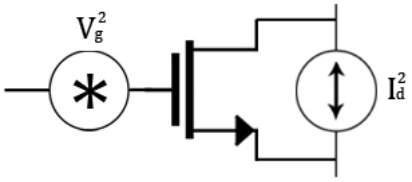

Figure B.1: Noise model for a MOS transistor. If frequency is low to moderate, the current source can be combined with the gate voltage source.

Where k is the Boltzmann constant and γ is a device constant, typically around 2/3. Note that for moderate frequencies, the thermal noise can be input-referred, leading to a simplified model where all noise sources are grouped at the gate and defined by:

$$V_i^2(f) = 4kT\left(\frac{2}{3}\right)\frac{1}{g_m} + \frac{K}{WLC_{ox}f} \tag{B.3}$$

Notice at low frequencies, the flicker noise dominates and is also governed by the device size. Hence, it is typically advantageous not only for layout but also for noise to have large transistors.

Before moving to amplifier level noise analysis, the cascode topology will be considered. For low frequencies (tested through simulation for typical transistor sizes, this is usually < 10 MHz), the input-referred noise if given by:

$$V_{n,in}^2(f) = \frac{K}{W_1L_1C_{ox}f} + \frac{g_mR_L}{(g_mr_o)^3} * \frac{K}{W_2L_2C_{ox}f} \tag{B.4}$$

$$V_{n,in}^2(f) \approx \frac{K}{W_1L_1C_{ox}f} \tag{B.5}$$

Therefore, this result leads to the conclusion that at low frequencies, the cascodes noise behaviour is governed by the common-source transistor. In other words, by increasing the length of the transistor, the noise level decreases leading to the noise advantage of the series-stack.

B.2 Op Amp Level

When speaking about the noise performance of an entire amplifier, it is typically broken down into the individual stages. The first stage, being the differential pair exhibits noise due to all four transistors. Most of the analysis shown here has been taken from the Gray and Meyer textbook [17]. Evaluating each transistor separately, assuming PMOS input transistors and an NMOS current mirror, then applying superposition leads to the following flicker noise equation:

$$V_{n,in}^2(f) = \frac{2K_p}{W_p L_p C_{ox} f}\left(1 + \frac{K_n \mu_n L_p^2}{K_p \mu_p L_n^2}\right) \tag{B.6}$$

Again, this equation only holds if the frequency is moderately low. To ensure that the second term in this equation does not appear and make things complicated, the NMOS is chosen to be twice the length of the PMOS. Note that the square makes this a factor of 4 difference, which typically overcomes the differences in the mobility and flicker noise constant to ensure that this factor is less than unity.

It is convenient to make the noise mostly dependant on the input device, since it can then be sized larger if need be, while maintaining the independence of the NMOS current mirror. For those extra curious, the thermal noise of the differential pair is given by the following expression:

$$V_{nTh,in}^2(f) = 4kT\frac{4}{3\sqrt{2\mu_p C_{ox}(W/L)_p I_D}}\left(1 + \sqrt{\frac{\mu_n (W/L)_n}{\mu_p (W/L)_p}}\right) \tag{B.7}$$

Note that the analysis done here assumes that the noise is dominated by flicker for the frequencies of interest. At higher frequencies, the white noise caused by thermal effects dominates until device capacitances enter into the picture.

Appendix C

Series-stack 3L

The analysis of the series-stack within the text was fairly straight forward since assuming the modes of operation (triode or saturation) were evident. When additional devices are placed in series, it becomes a little more complicated to determine the operating state of the middle device. For example, consider figure C.1. It is safe to assume that the Q_1 is operating in triode, while Q_3 is in saturation. The problem comes in determining Q_2.

Essentially, there are only two options, Q_2 could be in triode or saturation. If Q_2 is in triode, the analysis is simple since one can use the result of the two device series-stack. In other words, there is a saturation and triode device connected in series, which is known to result in a single saturation device with double the length. Hence, the final result is two devices, where Q_2 and Q_3 are combined into one saturation device. This can then be analyzed by using the same simplification, resulting in a single saturation device where the length is $3L$.

If Q_2 is in saturation, the analysis becomes challenging. However, it can be shown that Q_2 is always in triode. Assume Q_3 is in saturation, the 'on' condition can be written:

$$V_i - V_2 > V_t \tag{C.1}$$

This puts a constraint on the V_2 node:

$$V_i - V_t > V_2 \tag{C.2}$$

Considering Q_2, the condition for saturation can be expressed:

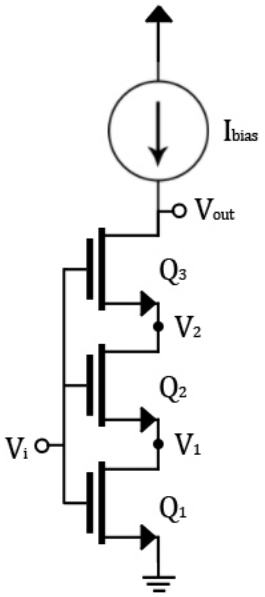

Figure C.1: A series-stack topology with three devices connect in series.

$$V_2 \geq V_i - V_t \tag{C.3}$$

Combining equation C.2 and C.3, a combined constraint can be derived for the V_2 node:

$$V_i - V_t > V_2 \geq V_i - V_t \tag{C.4}$$

Obviously, both of these cannot be true, but equation C.2 must be true to conduct current, meaning that Q_2 is not in saturation but in triode. It could be argued that Q_3 is in subthreshold, which would then leave the possibility that Q_2 is in saturation. However, in modern technologies, specifically below 100 nm resolution, the transition between subthreshold, triode and saturation is blurred. So, this entire analysis becomes difficult, and inaccurate.

As mentioned, it is safe to assume that the top most transistor of any series-stack will be in saturation. If this is true, then all subsequent transistors will have to be in triode; hence, any N number of L length series-stack devices will operate as a single saturation device with $N * L$ length.

Appendix D

Series-stack Small-signal Analysis

Although the DC analysis could be shown to yield the same current and voltage behaviour, it is useful to investigate the small-signal model in hopes of gaining further insight. First, the low-frequency model will be analyzed followed by the high-frequency model.

D.1 Low Frequency Behaviour

The low-frequency model is shown in figure D.1. Finding the transconductance can be found by doing KCL at the output:

$$i_{sc} = -g_m v_{gs} + v_1/r_o \tag{D.1}$$

$$i_{sc} = -g_m(v_{in} - v_1) + v_1/r_o \tag{D.2}$$

$$i_{sc} = -g_m(v_{in} - i_{sc}r_{ds}) + i_{sc}r_{ds}/r_o \tag{D.3}$$

Resulting in,

$$G_m = i_{sc}/v_{in} = \frac{-g_m}{1 + g_m r_{ds} + r_{ds}/r_o} \approx \frac{-g_m}{1 + g_m r_{ds}} \tag{D.4}$$

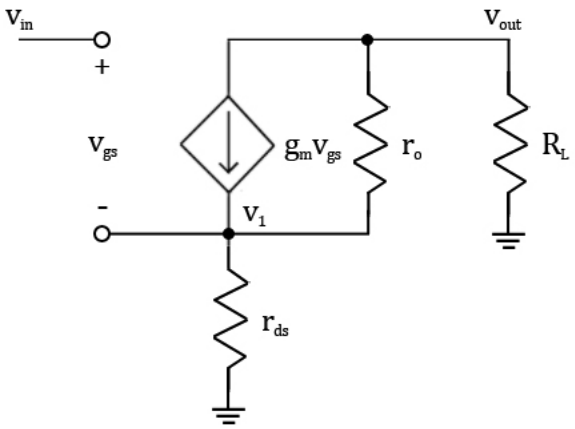

Figure D.1: Low-frequency small-signal model of the series-stack.

To make the analysis more insightful, the transconductance could be further approximated to $-1/r_{ds}$. To evaluate this approximation, it would be useful to calculate an expected value for r_{ds}.

Finding the overall gain also requires an output resistance value. It is clear that the small-signal model is equivalent to the output resistance of a cascode, leading to the following resistance:

$$R_{out} = r_{ds} + r_o + g_m r_o r_{ds} \tag{D.5}$$

Combining D.4 and D.5 leads to the following overall low-frequency gain equation shown in D.7. Although it may be satisfying that the result is the same as a single transistor, note that r_o is different than a similar saturation device with double the length, while g_m remains the same. So, it is expected that the series-stack would exhibit slightly lower gain, which is found in the results section of this work.

$$A_o = \left(\frac{-g_m}{1 + g_m r_{ds} + r_{ds}/r_o} \right) (r_o + r_{ds} + g_m r_o r_{ds}) \tag{D.6}$$

$$A_o = \left(\frac{-1}{r_{ds}} \right) g_m r_{ds} r_o = -g_m r_o \tag{D.7}$$

D.2 High Frequency Behaviour

A simplified high-frequency small-signal model is shown in figure D.2. It is simplified since some parasitic capacitors have been grouped into C_m, C_x and C_L. Exact analysis will be left for future work, but an open-circuit time constant analysis can be readily employed to gain insight. Note that this analysis is similar to that of a cascode.

Starting with C_L:

$$R_{CL} = (r_o + r_{ds} + g_m r_o r_{ds})||R_L = R_o||R_L \tag{D.8}$$

C_x and C_m experience the same resistance:

$$R_{Cx} = R_{Cm} = r_{ds}|| \left(\frac{r_o + R_L}{1 + g_m r_o} \right) = r_{ds}||R_i \tag{D.9}$$

It can also be shown that the resistance experienced by C_{gd} is that same as C_L, or $R_{Cgd} = R_{CL}$. Combining these results leads to the following time constant equation:

$$\tau_H = (R_o||R_L)[C_L + C_{gd}] + (r_{ds}||R_i)[C_x + C_m] \tag{D.10}$$

At this point, approximations are typically made to gain insight. However, it is clear that there is a dominate factor, involving C_L, which is similar to that of a single device. As expected, there is an additional pole due to the v_1 node. Although significant, it is not dominant, and hence should not drastically change the AC performance. This is reinforced by the results found in this work, where the series-stack is noticeably slower, but not by a significant amount.

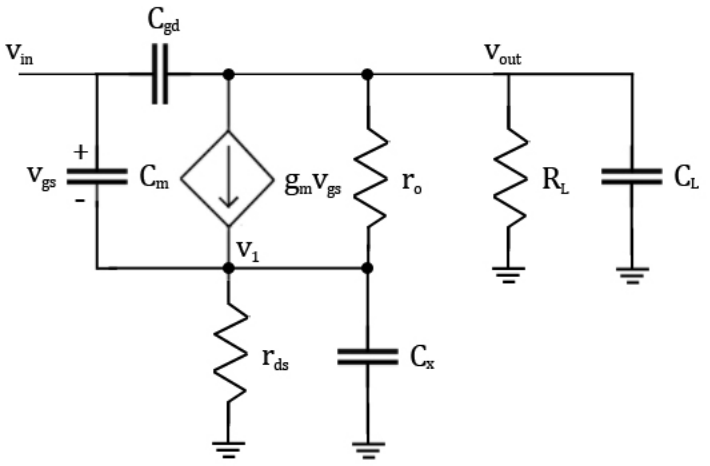

Figure D.2: High-frequency small-signal model of the series-stack.

Appendix E

Pole Optimization

As promised, an alternative approach for optimizing the second pole location will be presented. It is important to realize that much of this analysis is mute if there is a noise specifications, but since in this design there is none, such analysis is needed.

A problem manifests when trying to optimize the second pole equation 2.15, due to the transconductance stage being dependant on the scaling of the device, which also affect the parasitic capacitance. Consider the transconductance below:

$$g_{m2} = \sqrt{2k_n \left(\frac{W_\mu}{L}\right) I_\mu} * k = K_n * k \tag{E.1}$$

Where K_n is a constant dependant on device parameters and k is the second stage scaling factor. The parasitic capacitance, assumed to be dominated by C_{gs} can also be written:

$$C_p = kW_\mu * C = k * C' \tag{E.2}$$

Realizing that C' is a technology parameter, the second pole equation can be rewritten as follows:

$$\omega_{p2} = \frac{K_n * \left(\frac{C_p}{C'}\right) * C_C}{C_C C_C + C_C C_P + C_L C_P} \tag{E.3}$$

$$\omega_{p2} = \frac{K * C_p * C_C}{C_C C_L + C_C C_P + C_L C_P} \tag{E.4}$$

To plot the equation, C_C or C_P must be known. It is possible to estimate C_C based on the previous g_{m2} equation by setting the derivative to unity. Note that this is a significant assumption since the author is simply picking a slope of unity and deeming this the optimum point (to a certain extent it is since the designer being to get less than what is being put in), but one must choose some path otherwise there are simply too many. Doing some arithmetic:

$$\frac{\omega_{p2}}{g_{m2}} = \frac{C_C}{C_C C_L + C_C C_P + C_L C_P} \tag{E.5}$$

$$\frac{\delta \omega_{p2}/g_{m2}}{\delta C_C} = \frac{1}{C_C C_L + C_C C_P + C_L C_P} - \frac{C_C(C_L + C_P)}{(C_C C_L + C_C C_P + C_L C_P)^2} \tag{E.6}$$

$$\frac{\delta \omega_{p2}/g_{m2}}{\delta C_C} = \frac{C_L C_P}{(C_C C_L + C_C C_P + C_L C_P)^2} = 1 \tag{E.7}$$

$$C_C = \frac{\sqrt{C_L C_P} - C_L C_P}{C_L + C_P} \tag{E.8}$$

This derived equation can now be used to determine C_C for the previous K, equation E.4, relationship. This leads to an equation that can now be plotted, shown in figure E.1. The map between C_C and C_P is shown in figure E.2.

$$\frac{\omega_{p2}}{K} = \frac{C_p * C_C}{C_C C_L + C_C C_P + C_L C_P} \tag{E.9}$$

After examining the second pole frequency, it is clear that an optimum arises. The range of C_P from 0.2 to 0.5 seems reasonable to take advantage of the optimum. One can then map these values using the C_C versus C_P plot, shown in figure E.2, in order to determine the optimal value for the compensation capacitor, which is the goal in the first place. Then, choosing a C_C of about $0.2C_L$ covers a suitable range for the optimum this confirms what was found in the previous analysis.

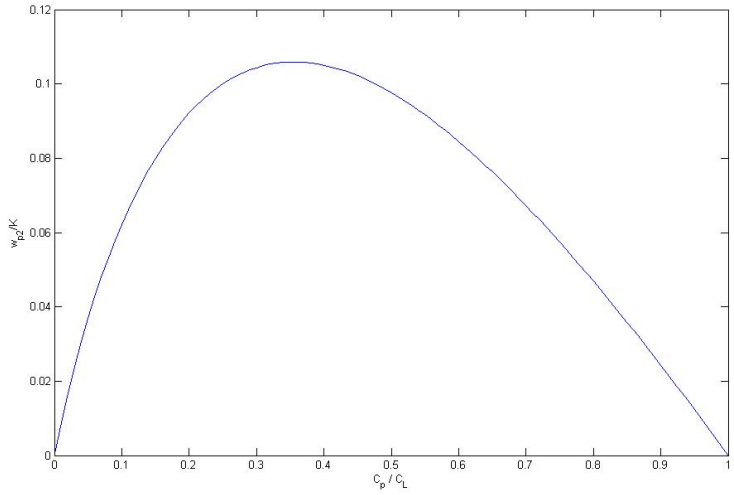

Figure E.1: An optimized second pole location plot.

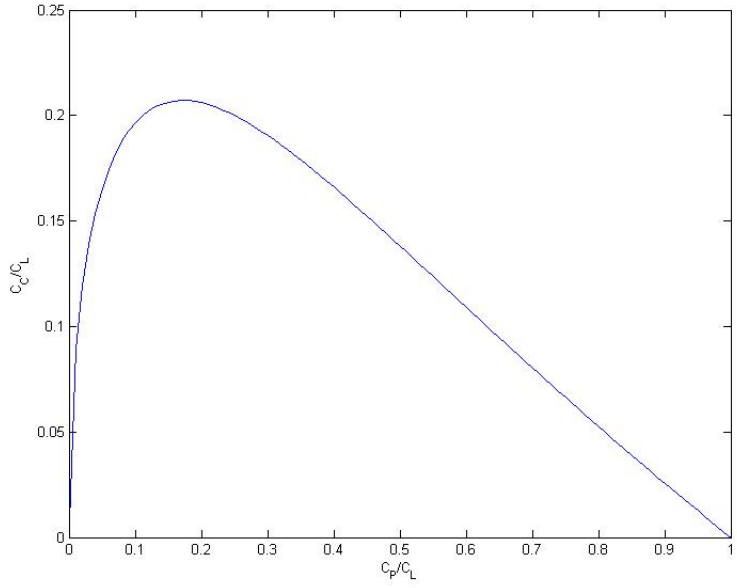

Figure E.2: The relationship map between Cc and Cp.

Appendix F

Op Amp Circuit Schematic

The complete circuit schematic of the bulk minimum length design is shown in figure F.1. Only the total widths are shown, even though all transistors were implemented using parallel techniques. Lengths are listed below widths.

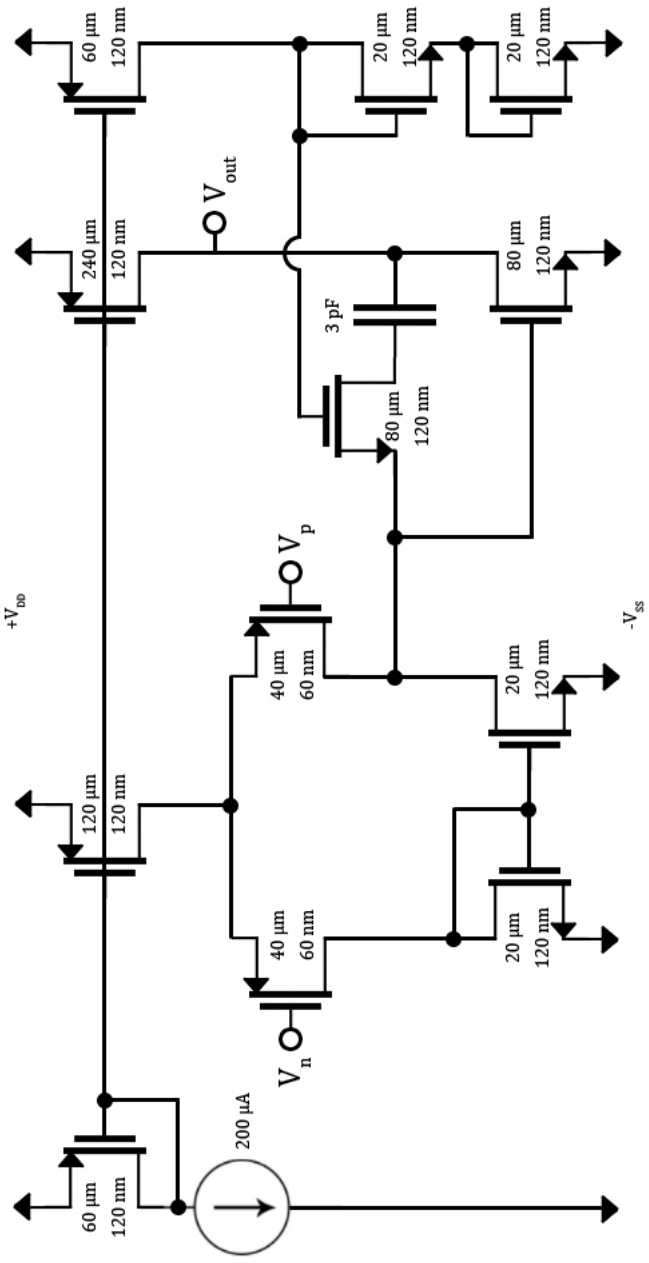

Figure F.1: Circuit schematic including all device sizes.

Appendix G

Testing Methodology

A quick review of the testing practices used to generate the simulation results will now be presented.

G.1 DC Gain, Unity-gain Frequency and Phase Margin

The three major specifications were measured using the same analysis, namely stability. The circuit topology is shown in figure G.1. An instance named "iprobe" was used for the analysis; it is a circuit component that acts as a short at DC but an open circuit at AC. Such an instance allows the bias to be properly set, while being able to break to loop and measure the loop gain as a function of frequency. Additionally, since the feedback topology is that of unity-gain, the loop gain directly reflects the overall gain of the amplifier ($\beta = 1$). It is also useful to characterize the amplifier in a unity-gain configure since this extracts the worst-case phase margin.

G.2 Settling Time

To plot the settling time, the amplifier was run in unity-gain configuration as shown in figure G.2. In a similar fashion to the phase margin measurement, the circuit topology uses a unity-gain configuration since this will yield the worst case (slowest) settling time. Note that the settling time was measured from the start of the signal shift to when the

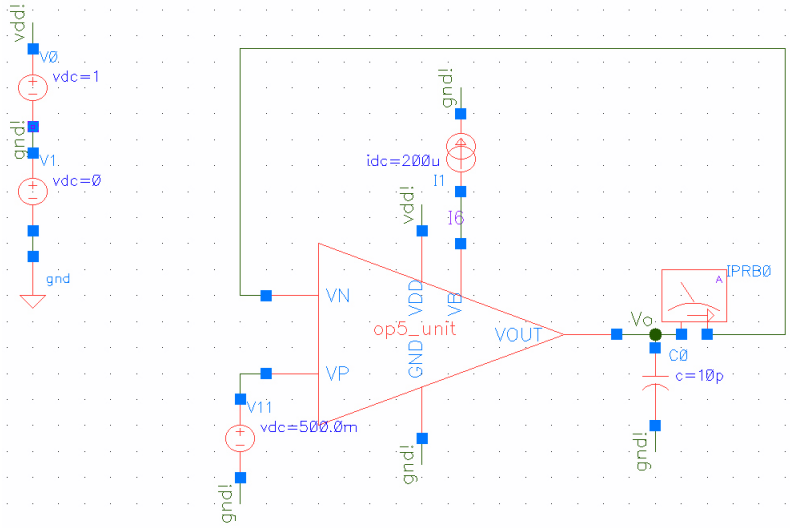

Figure G.1: Schematic setup for testing gain, speed and phase margin.

output signal reaches within 5% of the input signal. The settling time was also measured with the $10pF$ capacitance in order to reflect the eventual testing environment.

G.3 DC Parameters and Offset

This topology is used to extract DC parameters, as shown in figure G.3, and allows parameters such as output swing and offset to be calculated all from one simulation. The offset voltage is defined as the voltage needed to be applied to the positive terminal of the amplifier so that the output is exactly between the power rails. This was measured by simply sweeping the differential input voltages and measuring the output voltage. In addition, the simulation was run on slow-slow and fast-fast model parameters, the resulting difference between the measured offset voltage was called the "systematic offset variance". Notice that analysis involving Monte Carlo simulations and random offset can also be done; however, there were problems in the model libraries that did not allow for this. It is assumed that the systematic offset variance captures a similar understanding to the random offset measurement.

The output swing could be extracted by plotting the gain versus the output voltage. It was then defined that the limit of the output was when the gain dropped by 6 dB of its maximum value. Using this, the voltages were defined, and taking the difference led to the

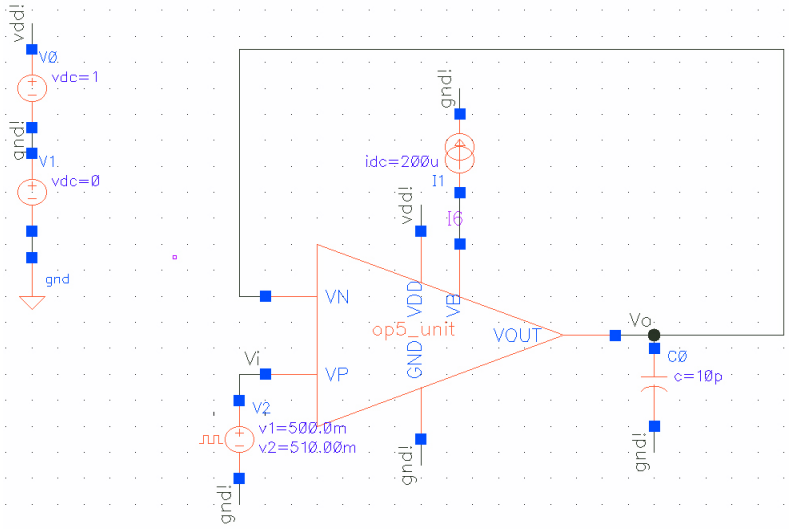

Figure G.2: Schematic setup to test transient behaviour.

output swing.

G.4 Common-mode Rejection Ratio

The setup used to test the CMRR is shown in figure G.4. A simple transfer function analysis (xf) was performed, and both the common-mode and differential-mode gains were plotted - the CMRR was taken as the ratio.

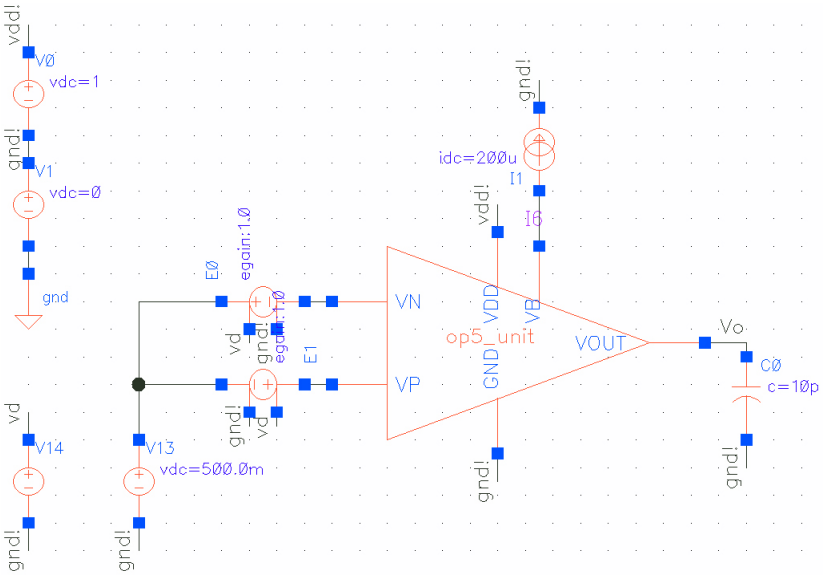

Figure G.3: General DC schematic setup for testing offset behaviour.

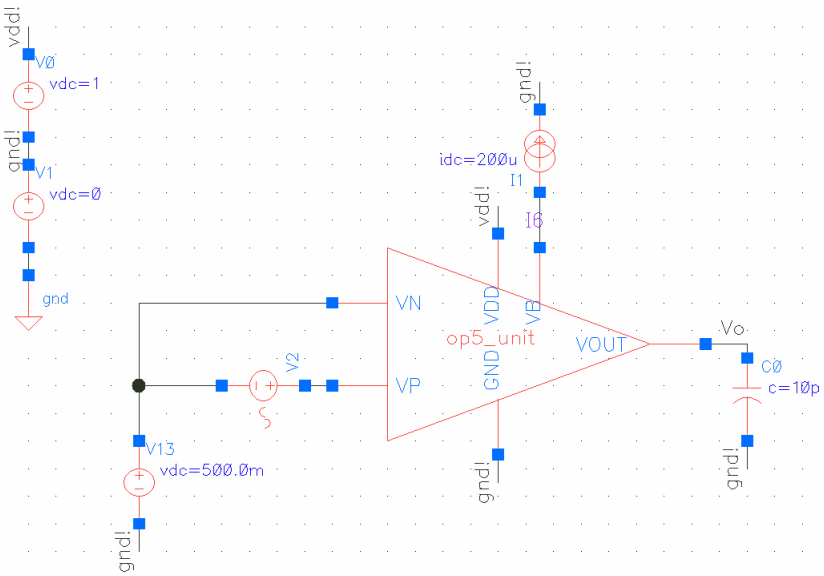

Figure G.4: Schematic setup for testing CMRR performance.

References

[1] A. S. Sedra and K. C. Smith, *Microelectronic Circuits*. New York: Oxford University Press, sixth ed., 2010.

[2] H. S. Black, "Inventing the negative feedback amplifier," *IEEE Spectrum*, vol. 14, pp. 54–60, 1977.

[3] H. Nyquist, "Regeneration theory," *Bell Systems Technical Journal*, vol. 11, pp. 126–147, 1932.

[4] J. E. Solomon, "The monolithic op amp, a tutorial study," *IEEE Journal of Solid-State Circuits*, vol. SC-9, pp. 314–332, 1974.

[5] D. A. Hodges, P. R. Gray, and R. W. Broderson, "Potential of mos technologies for analog integrated circuits," *IEEE Journal of Solid-State Circuits*, pp. 285–293, 1978.

[6] Y. P. Tsividis, "Design considerations in single-channel mos analog circuit - a tutorial," *IEEE Journal of Solid-State Circuits*, pp. 383–391, 1978.

[7] A. D. Blumlein, "Improvements in or relating to thermionic valve amplifying arrangements," *British Patent*, vol. 482,740, 1936.

[8] D. Senderowicz, D. A. Hodges, and P. R. Gray, "A high-performance nmos operational amplifier," *IEEE Journal of Solid-State Circuits*, vol. SC-13, pp. 760–768, 1978.

[9] P. Horowitz and W. Hill, *The Art of Electronics*. New York, New York: Cambridge University Press, second ed., 1989.

[10] P. R. Gray and R. G. Meyer, "Mos operational amplifier design — a tutorial overview," *IEEE Journal of Solid-State Circuits*, vol. SC-17, no. 6, pp. 969–982, 1982.

[11] T. C. Carusone, D. A. Johns, and K. W. Martin, *Analog Integrated Circuit Design*. Hoboken, New Jersey: John Wiley and Sons, second ed., 2012.

[12] B. Y. Kamath, R. G. Meyer, and P. R. Gray, "Relationship between frequency response and settling time of operational amplifiers," *IEEE Journal of Solid-State Circuits*, vol. SC-9, no. 6, pp. 347–352, 1974.

[13] F. Silveira, D. Flandre, and P. G. A. Jespers, "A gm/id based methodology for the design of cmos analog circuits and its application to the synthesis of a silicon-on-insulator micropower ota," *IEEE Journal of Solid-State Circuits*, vol. 31, pp. 1314–1319, 1996.

[14] B. E. Boser, "Analog design using gm/id and ft metrics." Online Document - http://www.eecs.berkeley.edu/ boser/presentations/2011-12, 2011.

[15] D. Hisamoto, W.-C. Lee, J. Kedzierski, H. Takeuchi, K. Asano, C. Kuo, E. Anderson, T.-J. King, J. Bokor, and C. Hu, "Finfet - a self-aligned double-gate mosfet scalable to 20nm," *IEEE Transaction on Electron Devices*, vol. 47, pp. 2320–2326, 2000.

[16] P. Mishra, A. Muttreja, and N. K. Jha, *FinFET Circuit Design*. New York: Springer, first ed., 2011.

[17] P. R. Gray, P. J. Hurst, S. H. Lewis, and R. G. Meyer, *Analysis and Design of Analog Integrated Circuits*. Hoboken, New Jersey: John Wiley and Sons, fifth ed., 2009.